2005-01-01

Onset and Ornament Detection and Music Transcription for Monophonic Traditional Irish Music

Aileen Kelleher
Technological University Dublin

Onset and Ornament Detection and Music Transcription for Monophonic Traditional Irish Music

MPhil Thesis by Aileen Kelleher

Faculty of Engineering

Dublin Institute of Technology

2005

Supervisors: Dr. Derry Fitzgerald

Dr. Eugene Coyle

Abstract

To date, much has been achieved in the areas of onset detection and music transcription although both still remain unsolved problems, particularly in the case of polyphonic music. This research focuses on detection of note onsets and pitches in monophonic music of three of the more popular instruments used by traditional Irish musicians. An attempt is also made at transcribing ornamentation, notes of extremely short duration, at most a fifth the length of a regular note. Ornamentation is a very important feature of this style of music, and its detection has not been previously attempted.

A thorough review of current onset detectors and music transcription systems was carried out. Various different approaches to solving the problem were encountered and each was assessed for its suitability for use in the proposed system. These techniques included the Short Time Fourier Transform, Autocorrelation and Wavelets.

By combining elements used in previous onset detectors, a hybrid system that detects note onsets and pitches in monophonic traditional Irish music has been implemented. The notes detected also include the most common types of ornamentation played by the fiddle, flute and tin whistle.

The proposed system used a Short Time Fourier Transform based sub-band technique, combined with an automatic threshold approximation to detect the note and ornamentation onsets. These onsets were then transcribed into the correct music notation. This system has been tested on a database of real recorded fiddle, flute and tin whistle tunes and good results have been achieved, particularly in the case of regular note onsets and pitches. The results for ornament detection, while not as good as those for regular onsets, is the first attempt at such detection and represents a good starting point for future research in ornament detection.

Acknowledgements

First of all I would like to thank my principal supervisor Dr. Derry Fitzgerald for all of his encouragement, guidance, support and patience without which this thesis would not be what it is today. Sincere thanks also to Dr. Eugene Coyle for his constant help and support.

Thanks to Matt Cranitch for informing me of this research project, and for the use of his fiddle tunes. Thanks to Dr. Bob Lawlor for his helpful discussions.

I would also like to thank my DiTME team members David Dorran, Mikel Gainza, Dan Barry and Jane Charles for their help, discussions and most of all their friendship along the way. Thanks also to Charlie Pritchard, Catriona Carthy and all in the DMC.

Thanks to all of my great friends and housemates for providing plenty of entertaining distractions and for always managing to cheer me up on days when things weren't going so well.

Most importantly thanks to my wonderful family, including Paula, Ali and particularly my parents Dan and Marian for their endless patience, friendship and support. I would not have accomplished this without their constant encouragement.

Table of Contents

Table of Figures

Table of Results

1. Introduction

Instrumental music transcription is the process of converting a musical sound signal into music notation, which a musician is able to read and play. Completing this task by hand is a long and tedious process, particularly in the case of polyphonic sounds where more than one note is played simultaneously. An ideal solution would be a software package, which converts a musical sound signal into a musical score at the click of a mouse. Developing such a system is not as simple as it sounds and the difficulty of transcribing music into notation depends on the complexity of the sounds with which it is presented. Clearly, the simplest sound to process would be electronically generated sinusoids whose starting and stopping times were controlled by computer. Real world musical signals are far more complex however, with even monophonic sounds presenting a range of challenges, whereas accurately transcribing a full symphony would be extremely difficult.

Transcribing a passage of music involves determination of the pitch [Howard '00] of each note that was played and the starting and stopping times associated with each of these notes. When a note is played, the frequency that is heard is known as its pitch. This note pitch typically consists of a fundamental frequency, also known as the first harmonic, and a number of related harmonics, which make up the harmonic series of the note. These harmonics occur at integer multiples of the fundamental. The second harmonic is located at twice the frequency of the fundamental, i.e. exactly an octave higher, the frequency of the third harmonic is three times the fundamental and so on. An example of the frequency spectrum of a note and its harmonics is shown in Figure 1.1. The fundamental frequency, $f0$, of the note is on the left with its harmonics to right.

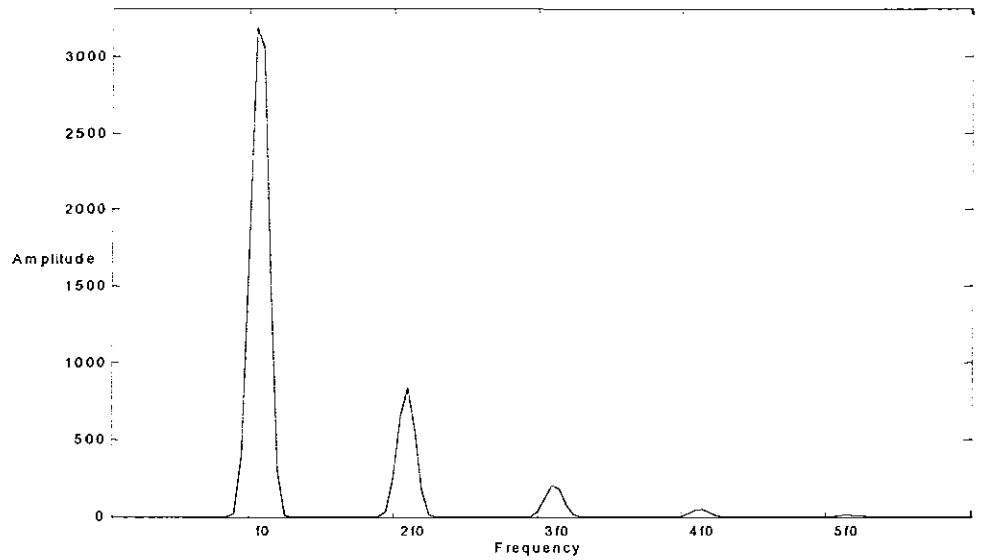

Figure 1.1: Note harmonics

The fundamental frequency is synonymous with the pitch of a note. Therefore pitch detection is the process of identifying the fundamental frequency of a note. In the above example, the fundamental is easy to spot since its amplitude is considerably greater than those of the harmonics, however, this is not always the case. There also exists a phenomenon known as the missing fundamental where only the second and greater harmonics are present in the note spectrum. In the case of polyphonic signals, when two notes are played simultaneously an octave apart, the fundamental of the upper note has the same frequency as the second harmonic of the lower note. Inharmonicities occur when the harmonic series of a note differs from the integral relationship f_0, $2f_0$, $3f_0$, etc. shown above in Figure 1.1, and individual harmonics waver slightly from multiples of the fundamental. For example a piano is inharmonic and its timbre is partly due to these inharmonicities. These examples give an indication of why pitch detection is difficult.

A number of different ways of determining the pitch of a note are outlined in the literature review. These techniques include both time and frequency domain techniques with the latter proving the more popular. Examples are Zero-Crossings [Cooper '94], Autocorrelation [Brown '91], [Bello '00], [Monti '00], Short Time Fourier Transform [Martin '96a], [Klapuri '00, '01], [Goto '00], The Constant Q Transform [Blankertz] and Wavelets [Chan '00]. Initial attempts at transcription

systems include that by [Piszczalski '77] in the frequency domain, others [Brown '91] have used time domain methods. [Martin '96a], [Klapuri '98] and [Goto '00] are among those who have attempted polyphonic transcription. Developments in the area of onset detection have been beneficial in determining the starting time of a note. The first onset detectors considered the musical signal as a whole [Chafe '85], while others [Scheirer '98], [Klapuri '99] expanded on this approach by dividing the signal into a number of frequency bands.

The main objective of this project is to accurately transcribe monophonic traditional Irish folk music. In addition to the problems of accurately describing the melody, a major challenge will be the degree of accuracy to which the ornamentation can be represented. Ornamentation consists of a note, or notes, of extremely short duration used to embellish a passage of music. It is an essential feature of this style of music, adding greatly to its character and charm and is discussed in detail in Chapter 2.

Traditional Irish tunes are generally played quite quickly and transcribing them manually can be difficult. A tool that would carry out this task automatically would be extremely beneficial, particularly as a learning aid for beginners. While a transcription system would be useful to an established musician, they may not require a device that detects ornamentation, as they would intuitively know when it is required. However an average beginner would certainly not have this skill. Ornamentation detection is not something that has been attempted previously and it would be advantageous for musicians to be able to transcribe a tune played by one of their peers to learn how ornamentation should be played. It would also be useful for a musician to transcribe their own work and compare to the correct notation in order to see where mistakes are being made.

The remainder of the thesis is comprised of the following chapters:

2. Traditional Irish Music.
   This chapter provides an introduction into the history of traditional Irish music. Ornamentation is described in detail and the different types are explained. Classical music is used as an aid to do this. Each of the instruments that will be used to test the system are introduced and their note ranges defined.

3. Literature Review

This chapter gives an overview of existing techniques that are being used in an attempt to solve the problems of onset detection and music transcription. It is divided into two sections, music transcription and onset detection. The reason for this is because onset detectors are used in other applications besides music transcription, such as instrument separation and time scaling, and is an area of research in its own right.

4. Proposed Approach

This chapter describes the proposed onset and ornament detection and music transcription system for monophonic traditional Irish music. It goes through various different techniques currently being used and explains why the proposed method was chosen. Results for the system are also presented in this chapter.

5. Conclusions and Further Work

This chapter comprises of comments on the strengths and weaknesses of the system. It also contains suggestions for additions and improvements to the system.

6. References

Contains the list of publications referenced in the text.

Appendix A

Contains an index of where to find each of the tunes used to test the system.

2. Traditional Irish Music

Traditional Irish music is one of the richest treasuries of folk music in the world. It belongs to an oral tradition; therefore, much of it has already been lost. It is now a sophisticated listening music and no longer simply an accompaniment for dancing. Since the late 1960s, there has been a radical change in the traditional Irish music scene. The commercial life of this style of music has mushroomed, bringing with it a huge growth in music tourism. Sony and JVC have invested in promoting the music abroad and there are approximately one thousand specialised albums available. There is considerable commercial interest in the native music of an island of approximately five million people [McGettrick '99].

The significance of traditional Irish music is clear from a number of important studies that have been done. One of these [Fleischmann '98] was begun by Professor Aloys Fleischmann in 1950. It was finally completed at the Irish World Music Centre, University of Limerick under the direction of Professor Mícheál Ó Súilleabháin in 1998. It is a commented compilation of all traditional tunes recorded in Irish manuscript and printed collections – including related material in Scottish, English and Welsh collections, and in eighteenth-century ballad operas - from the very first c. 1583 up to George Petrie's *Ancient Music of Ireland* of 1855. It is the largest publication of any kind on Irish traditional music and is the result of over forty years of research. It contains almost seven thousand tunes, presented in chronological order, with notes and an analysis of tonality and structure. There is a comprehensive set of indexes, including an index of incipits. An incipit is the first part of a tune, encoded into a series of letters or digits. An example is the do-re-mi format where do represents the root of the major scale. Note lengths and sharps or flats are not expressed. This vast corpus of folk music was largely inaccessible before this publication since all but a few of the sources are out of print. Bringing the sources together for the first time makes material available, which allows a mapping of evolution and change in traditional music in Ireland up to the middle of the Nineteenth Century and brings to light many traditional Irish airs that had fallen out of use in the living tradition.

Many traditional Irish tunes have been lost over the centuries, mainly because they have not been written down or recorded. Often, many different versions of the one tune can exist, depending on factors such as geographical location, and musicians can add their own personal touches as they play. It is clear that the development of a music transcription system capable of accurately transcribing monophonic traditional Irish folk music would be a significant help in preserving this tradition. It would enable session musicians to transcribe their recordings quickly and easily into musical notation where they could be preserved indefinitely.

2.1 Ornamentation

Traditional Irish music contains a great deal of ornamentation. It is extremely important to the character and style of the music since it is what gives the tunes their depth and richness. Without it, the faster tunes tend to lose some of their distinct liveliness; the quality that drives people to instinctively tap their foot as they listen; and ballads become less melodic. These differences are particularly noticeable to the trained ear. Ornamentation is not unique to traditional Irish music as it is present in most types of folk music and is featured in Baroque music. The significance of ornaments was summed up by [C.P.E. Bach 1753]:

> *"It is not likely that anybody could question the necessity of ornaments. They are found everywhere in music, and are not only useful, but indispensable. They connect the notes; they give them life. They emphasise them, and besides giving accent and meaning, they render them grateful; they illustrate the sentiments, be they sad or merry, and take an important part in the general effect. They give to the player an opportunity to show off his technical skill and powers of expression. A mediocre composition can be made attractive by their aid, and the best melody without them may seem obscure and meaningless."*

In other words, ornaments are used by folk and classical music to embellish a note or passage of music. Although this is where the main similarities between the two genres end, classical music will be used as a guide to help define the characteristics of ornamentation in traditional Irish music. In classical music, an ornament is considered a note or group of notes in their own right, whereas in traditional Irish music it is not

considered a note; merely an alteration or embellishment belonging to a parent note. These alterations and embellishments are created mainly through the use of special articulations and inflections, not through the addition of extra, ornamental or grace notes [Larsen '03].

The ornamentation present in traditional Irish music today originated from the Irish bag piping tradition. The Uileann pipes are the modern Irish bagpipes; their closest ancestor was most likely the pastoral pipes. Unlike the uileann pipes, it was not possible to stop the flow of air through the pastoral pipes resulting in a constant stream of sound. The only way a player had of breaking this sound was by moving their fingers into different positions. Therefore, in order to separate notes of the same pitch played in succession a finger articulation was used. Today, the fiddle is the most widely used instrument in Irish traditional music. It is not a native instrument as it is the same as a normal violin but the style of playing is distinctly Irish [Cranitch '01]. Another instrument regularly heard in traditional Irish music is the tin whistle. It originated as far back as the third century A.D. [McCullough '87], although the earliest evidence of its use in traditional Irish music appears to be at the beginning of the 19th century [O' Farrell 1804]. It was not until the 1960s that the tin whistle was taken seriously in Ireland, it was previously considered as more of an introductory instrument. The whistle comes in a number of different sizes and keys with the most common being the small D whistle. The Irish flute is again non-native to Ireland. There are many different types; the 19th century classical wooden flute and modern instruments closely based upon it is the most favoured by Irish musicians. Other commonly used instruments include the accordion, guitar and banjo.

The proposed system is concerned with the music of the fiddle, flute and tin whistle. A database of music was formed using fiddle tunes from [Cranitch '01] and tin whistle and flute tunes from [Larsen '03]. These tunes included jigs, reels, slides, polkas and hornpipes and were studied to determine the characteristics of traditional Irish ornamentation. This database was found to contain approximately 1600 ornaments of various different types.

2.1.1 Types of ornaments

In classical music [Blood '02], ornaments are not necessarily written into the score although this is often the case. A performer is also free to add their own as a way of expressing their individual style of playing. They can be left out or added to a piece without fundamentally changing it. Examples of ornaments used in classical music are:

1. Grace Notes
2. Trills
3. Mordents
4. Turns

They are sometimes combined, for instance a grace note may be played with a mordent or a grace note with a trill. The most simple ornament is a grace note, which is a short passing note played immediately before a parent note. A grace note is notated as a small note and its duration is not specified, examples are shown below in figure 2.1. The two most important types are acciaccatura and appoggiatura. An acciaccatura, figure 2.1 (a), is a crushed dissonant note of the shortest possible duration played either on the beat or just before the parent note and immediately released. An appoggiatura is an ornament note that is usually one step above or below the note it precedes. It begins either just before the beat, borrowing the time it occupies from the note preceding it, figure 2.1 (b), or on the beat borrowing the time it occupies from the parent note, figure 2.1 (c).

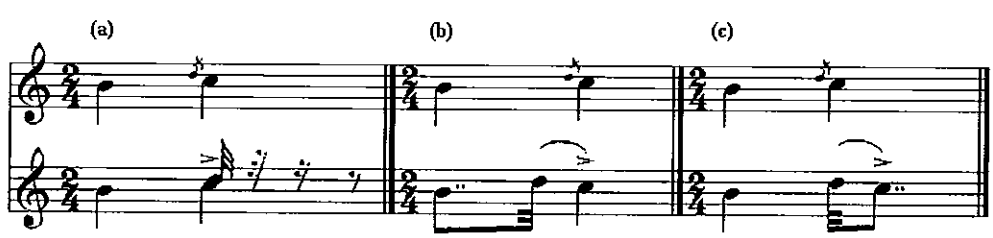

Figure 2.1: Examples of grace notes [Blood '02]; (a) acciaccatura, (b) and (c) appoggiatura, the top line shows how they would be indicated in a musical score and the bottom line how they are played

A trill consists of two notes, it begins on the note above the principal note and finishes on the principal note. It consists of six or more individual notes, which are as short as the player is able to play, shorter than a grace note. There are two types of mordent; the 'lower' was used in music before the nineteenth century and the 'upper' in all music that followed. They both consist of three notes, beginning and ending with the principle note. In the case of the 'lower' mordent, the middle note is the note below the principle note and for the 'upper' mordent, the middle note is the note above the principle note. A turn consists of four notes, the note above the principle note, the principle note, note below the principle note concluding with the principle note. Each of the individual notes is generally longer than a grace note.

The most commonly used types of ornamentation in traditional Irish music are the:

1. Cut
2. Double Cut
3. Strike
4. Slide
5. Roll

In our database, the double cut was exclusive to the fiddle and the strike and slide to the flute and tin whistle. The cut and roll were used by all three instruments. The cut, strike and slide are single note ornaments, the double cut contains two cuts and the roll is a multi-note ornament, which involves using a combination of cuts and strikes. The cut could be considered as the equivalent to a grace note. However, it is not considered to be a note independent of the parent note to which it is attached, as is the case for a grace note. Nor is it considered to have a definite pitch or duration and it is played on the beat of the parent note. It is at the musicians' discretion what note is used to 'cut' the parent note as all players have their own style.

The fundamental difference between a cut and a strike is that a cut is pitched above its parent note and a strike below. Whereas the cut and strike are articulations that have an instantaneous effect, a slide is an inflection and is a continuing, moving alteration of a note's pitch [Larsen '03]. There are two types of slide, falling and rising, where the preceding note pitches are above and below the parent note's pitch respectively. There are also two types of roll, a short roll and a long roll. These are similar to a trill

in classical music. The short roll consists of cut-note-strike-note and a long roll consists of a short roll preceded by a parent note. Examples of these ornaments can be seen in figure 2.2, which uses a classical interpretation to define them, i.e. the ornaments are depicted as notes. This figure is merely an aid to explaining what each ornament consists of and is not the way they would be indicated within a tune. In the two books used in this study [Cranitch '01] and [Larsen '03], different symbols were used by both musicians to indicate when they should be played.

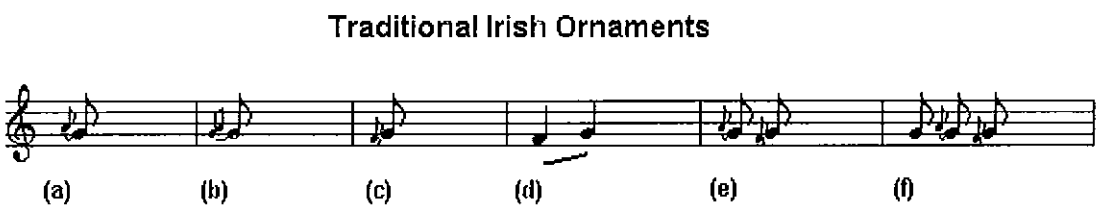

Figure 2.2: Different types of ornament s used in traditional Irish music; (a) the cut, (b) the double cut, (c) the strike, (d) a rising slide, (e) the short roll and (f) the long roll

In [Windsor '00], a study was carried out which investigated the execution of grace notes in a musical performance. By testing the relationship between different classifications of grace notes and their durations at different tempi, they attempted to show that the local structure of music has a systematic effect on both the durations of the grace notes and the relationship between these and local tempo measurements.

Five pianists each played a simple Beethoven piano piece, containing eleven grace notes, five times at nine different tempi resulting in forty-five performances. The time between the grace note and its parent note was measured as well as the time between the note preceding the grace note and the note following the parent note. The second measurement was divided by the number of eighth notes in that time span in order to calculate the local eighth note duration.

Of the eleven grace notes contained in the piece, three had an interval of a semi-tone with their parent note, six had a tone and two had an interval of nine semitones. It was found during the study that the grace notes with a large pitch interval were played for a longer duration than those with shorter pitch intervals. The grace note that was played for the shortest time had an interval of a semi-tone and was a black note with

its parent note being a white note. This lead to the conclusion that grace notes in different structural classifications will be played differently regardless of tempo; although they found that the major influence on grace note timing seems to be more stylistic.

Although the above study was carried out on classical music, it is also very relevant to traditional Irish music. In the study, they found that a players style has the most influence on the duration of a grace note. This is significant as it is a players style, which influences the interval between an ornament and its parent note. For example, different players use different notes with which to 'cut' the parent note. Some notes are also more difficult to cut than others, for example C#5 on a D tin whistle. A C#5 is played by simply blowing through a tin whistle without covering any holes, D5 the note above it is played by covering all holes and it is quite difficult to make the transition while including a cut. A large jump between the note preceding the ornament and its parent note can also cause a delay. All of this implies that an ornament will invariably have an effect on the duration of the parent note to which it belongs, with this effect being more noticeable in some circumstances than in others.

2.2 Note Range

To decide the note range of the proposed system, the tunes in the database were thoroughly examined. The range of notes used by traditional Irish musicians when playing the fiddle is somewhat less than that which would be used by a classical musician. In all only twenty-nine semi tones are used, those from G3 to B5. It is very unlikely that a note outside of this range would be played.

As previously mentioned, the small D whistle is the most common type played by traditional Irish musicians and is used in more than 80% of Irish tunes. This is a transposing instrument in that a note always sounds an octave higher than what was played from a score. An example would be when D4 is played from a score it is actually D5 that sounds. The range of this instrument is from D4 to B5. It is possible to play the next octave but notes of a higher pitch than B5 are shrill and unpleasant to the ear and are not played in traditional Irish music. The flute range is the same as that of the tin whistle. Consequently, the transcription system proposed later will deal with

the 29 semi-tones from G3 to B5 as this covered the range of each of the three instruments concerned.

2.3 Chapter Summary

This chapter has shown that traditional Irish music is a valuable corpus of music. A brief history of this style of music has been provided. Ornamentation has been described in detail in both classical and traditional Irish music with the more familiar structure of the former used as an aid to define the latter. The instruments that will be used to test the proposed onset and ornament detection and transcription system have been outlined along with each of their individual note ranges.

3. Literature Review

The main objectives of this literature review were:

- To obtain a better understanding of the problem that is automatic music transcription.
- To investigate previous and current methods being used in order to determine a suitable approach to accurately transcribe monophonic traditional Irish musical signals, including ornamentation.

The literature review is divided into two sections, 3.1 Music Transcription Systems and 3.2 Onset Detection. The first section deals with complete systems and pitch estimation whereas the second deals exclusively with onset detection. It was thought to be a more appropriate way to deal with the literature review, as onset detection is an area of work in its own right. Onset detectors are also used in applications other than music transcription, such as musical instrument separators [Virtanen '00] and time stretching [Duxbury '03a].

3.1 Music Transcription Systems

This section takes a look at existing methods of transcribing a musical signal. Initial attempts began back in the early seventies and although much progress has been made over the past few years, the problem still remains unsolved. There are two types of transcription, monophonic, where one note is played at a time and polyphonic, where two or more notes are played simultaneously. This review deals with both types. Both time-domain techniques such as autocorrelation and frequency domain techniques such as the Short Time Fourier Transform (STFT) have been used, with most opting for the latter.

[Knowlton '71] developed a system at the University of Utah using a hard-wired piano keyboard as the system input device; however, the system required specially modified musical instruments to produce the notation.

Initial attempts at an acoustic signal-based transcription system were described in [Piszczalski '77]. He developed an automatic music transcription system that focused

on the recorder and symphonic flute, although low-level analysis was also done on the piano and the violin. He described how a music transcription system consists of three distinct parts:

1. Detection of component frequencies and their starting and stopping times along with associated amplitude values.
2. Analysis of the frequency, amplitude and time information produced to determine the musical notes that are present.
3. Production of the final notation, determining the location of the measure bars, etc.

The musical signal was converted into a time frequency representation using the Short Time Fourier Transform (STFT) [Allen '77], which is derived from the Fourier Transform. The continuous Fourier transform is used to transform a continuous time-domain signal into its continuous frequency-domain equivalent by representing the signal as a set of sine waves of different frequencies. It is defined as follows:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft}\, dt \qquad (3.1)$$

where $x(t)$ is the continuous time-domain signal and $X(f)$ its frequency decomposition. The Discrete Fourier Transform (DFT) is used to calculate the frequency content of a discrete signal sequence and is defined as follows:

$$X(k) = \sum_{0}^{N-1} x(n) e^{-j2\pi nk/N} \qquad (3.2)$$

where $x(n)$ is the discrete time-domain signal and $X(k)$ its frequency content. Unlike the continuous Fourier transform, the DFT covers a finite time and frequency span. The original signal can be retrieved from $X(k)$ using the Inverse Discrete Fourier Transform (IDFT):

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N} \qquad (3.3)$$

The STFT is an analysis method that uses the DFT.

$$X(k,n) = \sum_{m=0}^{L-1} x(m+nH)w(m)e^{-j2\pi mk/N} \quad (3.4)$$

where $x(m)$ is the signal, k the frequency bin number, n the frame number, H the hop length and $w(m)$ the window of length L. The STFT involves dividing a time-domain signal of any length into shorter blocks, or frames, and applying the DFT to each of these blocks individually. The frame length must be less than or equal to the DFT length. A frame is constructed by multiplying the time-domain signal, $x(n)$, which is in this case an audio waveform, by a window, $w(n)$. This is illustrated below in figure 3.1 where the original signal (a) is multiplied by a windowing function (b) resulting in a section of the original signal (c), which is known as a frame. The DFT is applied to this frame to obtain its frequency content. The windowing function is then shifted to the right and the process is repeated for the length of the signal. The STFT output can be interpreted as a collection of uniform filter outputs.

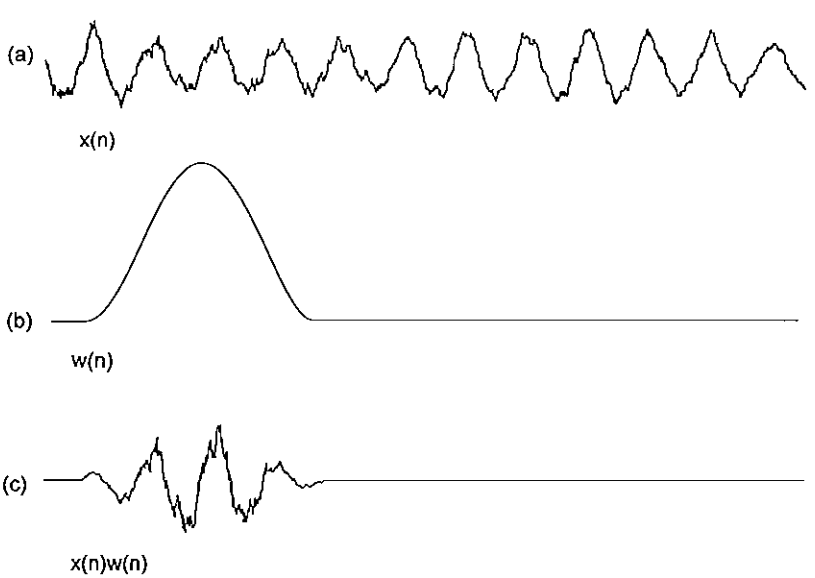

Figure 3.1: Creating a frame by multiplying a signal, x(n), by a window function, w(n)

The choice of analysis window [Harris '78] is important since good resolution in the frequency domain means a trade off in the time domain and vice versa, which is an

important factor to consider if the original audio waveform needs to be reconstructed from its frequency representation. The ideal window function for good frequency resolution would be of infinite length whereas for good time resolution it would be as short as possible. Obviously this is impossible so a compromise must be reached, and a window that fits somewhere between the ideal must be used. The simplest window function is the rectangular window. Examples of other windowing functions are Triangular, Hamming and Hanning. The responses of these windowing functions can be seen below in figure 3.2.

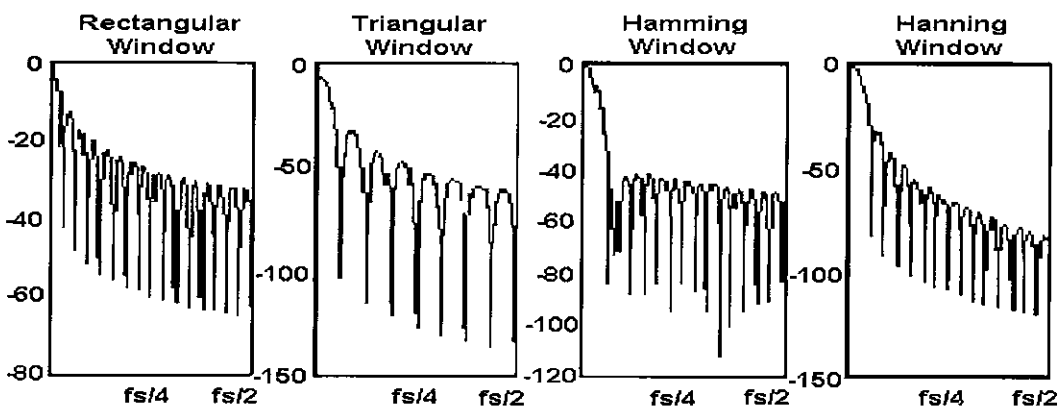

Figure 3.2: Responses of the Rectangular, Triangular, Hamming and Hanning windows

Once the STFT of the signal had been obtained, Piszczalski then extracted the frequency of strongest amplitude from each individual time section with each frequency and its associated amplitude value kept in time order. It was then possible to search for note boundaries by examining the peaks and troughs of this resulting strongest frequency line. It was assumed that once the amplitude dropped below the audible hearing threshold [Zwicker '99] a note offset had occurred and once the amplitude rose above this threshold a note onset had occurred.

However, Piszczalski found that for most musical instruments the strongest-frequency line did not contain enough information for complete identification of the musical notes and for more difficult note boundaries the joint time-frequency plane would need to be examined.

[Brown '91] used her previous findings on narrowed conventional and inverted autocorrelation to determine the fundamental frequency of musical signals produced by keyboard, wind, and string instruments. These results were compared to those obtained on the same sounds using conventional autocorrelation. By doing this, they were able to determine two things, whether the method of autocorrelation is well adapted to the problem of pitch tracking of musical signals and under what conditions narrowed autocorrelation is advantageous.

The autocorrelation of an N length sequence, $x(k)$, is calculated as follows:

$$r_{xx}(n) = \frac{1}{N} \sum_{k=0}^{N-n-1} x(k).x(k+n) \qquad (3.5)$$

where n is the lag and $x(n)$ is a time domain signal. An important property of musical sounds is that they usually have harmonic spectral components. This means that the second harmonic has a period equal to half that of the fundamental so its peaks in the autocorrelation function occur at T/2, 2(T/2), 3(T/2) ... n(T/2). The second peak of the second harmonic, then, will coincide with the first peak of the fundamental and similarly for other harmonics. Therefore a large peak corresponding to the sum of all spectral components should occur at the period of the fundamental (and all integer multiples of the period of the fundamental). It was this property that made the method of autocorrelation appear to be a good one to Brown for the pitch tracking of musical signals. Narrowed autocorrelation is calculated by including expressions in the autocorrelation function, which correspond to lags at $2n$, $3n$, etc. as well as the usual term with lag n. This results in an autocorrelation function with extremely narrow peaks.

Brown found that conventional autocorrelation returned the average of two notes in some transition regions whereas narrowed autocorrelation was better able to choose between the two. This mechanism was discussed in [Brown '89]. Two disadvantages of narrowed autocorrelation were that firstly, the longer analysis time meant less was known about the exact time for which the calculated note applied, i.e. there was the usual time/frequency trade off, and secondly the calculation was a little more expensive computationally as there was an extra N - 2 additions, where N is the number of terms.

[Cooper '94] developed a monophonic pitch-tracking algorithm, which was designed so that conventional instruments could be used to control musical processes in a way that was easily understood by the performer. His approach was to use the pattern and shape of the digitised sound wave to find its pitch. Using an oscilloscope you can usually see the points where a cycle repeats itself and, if the sampling frequency is known, determine the frequency of the wave. Cooper discussed the strategy of segmentation of the sound wave and a method of finding the shortest distance between two shape-similar repeating segments.

The upward zero crossing points of the sound wave were located. For a pure sine wave, the pitch could be calculated from the distance between two upward zero crossing points. However, the sound wave from a real musical instrument is more complex. The section between two zero crossing points was called a segment. A shape description was estimated by dividing a segment into eight equal sub-segments and taking the amplitude values of the first and last three of these sub-segments as six landmark points. These landmark points provided a simplified shape and were used to determine a similarity measure between the segments. The length of the largest segment was compared with all other segments and the distances and similarity ratios between them calculated. On locating the next similar segment, the frequency of that snapshot can then be calculated as follows:

$$frequency = \frac{f_s}{d} \tag{3.6}$$

where f_s is the sampling frequency and d is the distance between the two similar segments.

[Martin '96a] described a system that transcribes simple polyphonic music using a blackboard system. A blackboard system [Corkill '91] is based on the idea of a number of experts in various different fields congregated around a blackboard, working together to solve a problem. Each person will add his or her own expertise to help solve the problem when it is required. A blackboard system consists of the blackboard, knowledge sources and a scheduler. The blackboard would contain the various steps of the problem, the knowledge sources are equivalent to the experts and the scheduler decides who develops what and when. A knowledge source finds a

piece of relevant information on the board and works on it independently of the other knowledge sources. Knowledge sources can be added or removed as required without affecting the other knowledge sources.

The front end of the system obtained the time-frequency representation of the signal using the STFT. In parallel with this analysis, the short-time running energy of the signal was also measured by squaring and low pass filtering of the original signal. Any sharp rises that are detected in the running energy were interpreted as note onsets and this information was used to segment the time-frequency information into chunks, which represent the individual chords. The time-frequency analysis of each segment was averaged over time to give an average spectrum for the chord that was played. The energy peaks were picked out as they corresponded to harmonic tracks. The input for the blackboard system was now present, i.e., a list of harmonic tracks with their associated onset time, frequency and magnitude.

Martin was more interested in a simple front-end system rather than a robust one. Most of his work concentrated on developing the blackboard section. The blackboard workspace was divided into five levels, tracks, partials, notes, intervals and chords. Thirteen knowledge sources were described which each fell into one of three categories of knowledge, garbage collection, knowledge from physics and knowledge from musical practice.

The implemented system was capable of transcribing polyphonic synthesised piano performances in which no two notes were ever played simultaneously an octave apart and where all notes were in the limited range of B3 to A5. The front end of the system made the assumption that the notes in the chords are all played simultaneously and that the sounded notes did not modulate in pitch. Although this system suffered from some limitations, Martin considered it to be an important step towards solving the transcription problem.

[Martin '96b] uses a different front-end system than he used in [Martin '96a], based on autocorrelation as opposed to sinusoidal analysis. One of the difficulties experienced in [Martin '96a] was the problem of detecting octaves and it was hoped that this system would overcome that problem. Every note has harmonics at integer multiples of their fundamental frequency e.g., the note A4 will have its second

harmonic at A5. Therefore when two notes are played simultaneously an octave apart, the fundamental frequency of the higher note will occur at the same frequency as the second harmonic of the lower note. This leads to great difficulty in detecting the higher note.

The detection system was based on the log-lag correlogram of [Ellis '96]. The signal was split into frequency bands using forty gammatone filters, six per octave, spaced evenly in log frequency. This bank of filters models the basilar membrane mechanics of the human ear. Each band then underwent half-wave rectification followed by smoothing and onset enhancement, to model inner hair cell dynamics. The bands were then analysed using short-time autocorrelation. The three axes of the log-lag correlogram were filter channel frequency, lag (or inverse pitch) on a logarithmic scale and time. In Martin's implementation, a summary autocorrelation was obtained by normalizing each frequency / lag cell by the zero-lag energy in the same frequency band and averaging across the frequency bands. This computed the 'pitch percept'. The output from the correlogram becomes the input for the blackboard system.

The system contained only five knowledge sources and therefore much of the information contained in the summary autocorrelation was ignored. The results were encouraging and while the problem of octave detection was not overcome, it was thought that results could be improved through the addition of musical knowledge to the blackboard system. Overall, the results obtained for the sinusoidal modelling approach in [Martin '96a] were better.

[Sillem '98] described a system that would transcribe polyphonic music. The system would be required to handle multiple instruments playing simultaneously and was regarded as a feasibility study for a commercial package. The finished product would ideally be a black-box music transcription system that would have a microphone on one side and a MIDI [Rumsey '94] port on the other. The musical sounds played into the microphone would be converted into MIDI messages which could be recorded, notated or sent to a MIDI controlled instrument or synthesiser.

He identified a series of steps for converting digitised acoustic signals to MIDI note on/off messages i.e. a music transcription system:

1. Convert the wave data to a series of frequency domain snapshots.
2. Extract and characterise sharp peaks.
3. Link sharp features in consecutive frequency snapshots to form features in the time/frequency plot
4. Group simultaneous, harmonically related features in the time/frequency plot into notes.
5. Reject unwanted features and notes.
6. Identify the instrument playing the notes.
7. Output the notes as MIDI note on/off messages.

The project would be PC and windows based and use standard wave (.WAV) and MIDI files, which would allow maximum compatibility with existing sound recording and notation packages.

Various methods of frequency analysis were tested. Sillems preferred method was the FFT although frequency resolution was poor at low frequencies and unnecessarily good at high frequencies. Increasing the window length increased the frequency resolution. The other methods discussed were the maximum entropy method / all poles model, digital filtering and wavelets [Kaiser '99].

The maximum entropy method chooses a model, which is consistent with the facts it is presented with but is otherwise as uniform as possible i.e. it models all that is known while assuming nothing about the unknown. This method was mathematically well suited to modelling sharp features in a power spectrum as the poles in the continuous power spectral density function produced a set of very high, sharp peaks. However, the peaks wandered around unpredictably and sometimes split to produce two peaks where there should only be one. Sillem found that this caused unreliability.

For the digital filtering method, a bank of IIR filters was used, each tuned to a different note. This gave poor results with an unacceptable trade-off between time and frequency localisation. Improvements were seen when the order of the filter was increased, but this lead to an excessive computational load.

While the Fourier Transform represents a signal as a set of sine waves of different frequencies, the Wavelet Transform represents a signal as a set of shifted and scaled

versions of an original wavelet. The Morlet, Meyer and Mexican hat wavelet functions can be seen below in figure 3.3. A sinusoid has an infinite duration and is smooth and regular. A wavelet is a waveform of limited duration and is suited to detecting abrupt changes and beginnings and ends of events, features that are common in musical signals.

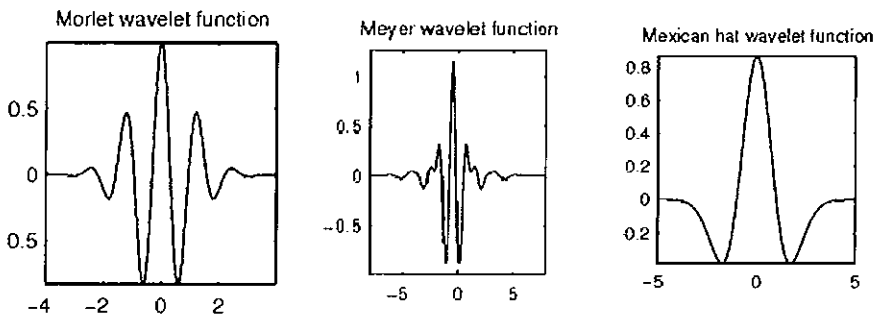

Figure 3.3: Morlet, Meyer and Mexican Hat wavelet functions

The continuous wavelet transform, C, of a signal, $f(t)$, is defined as follows:

$$C(scale, position) = \int_{-\infty}^{\infty} f(t)\psi(scale, position, t)dt \qquad (3.7)$$

where ψ is the wavelet function. An example of a scaled wavelet is shown in figure 3.4 where (a) is the original wavelet. A high scale, like that used in (b) would be used to analyse low frequencies and a low scale like in (c) is used to analyse high frequencies where more detail is required. The scaled wavelets are moved along the signal and correlated with it.

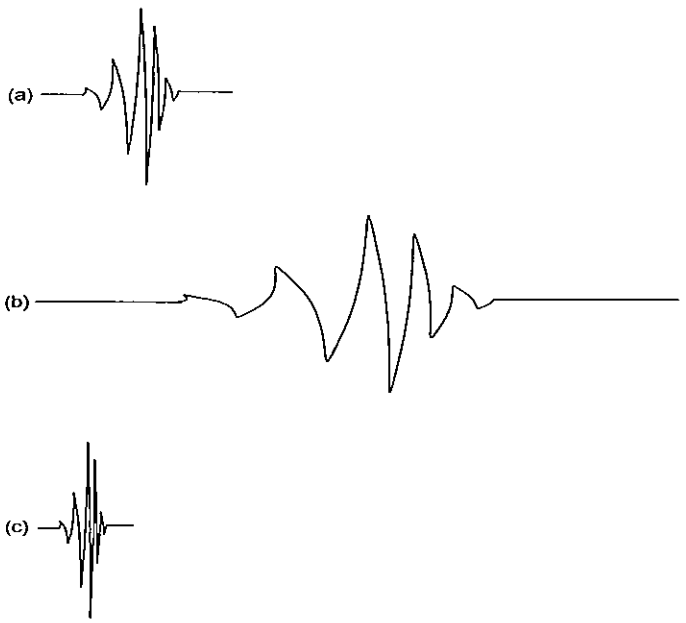

Figure 3.4: Examples of an original wavelet (a), a wavelet with a high scale (b) and a low scale (c)

Sillem found Morlet wavelets to be computationally similar to the "slow" Fourier Transform and not particularly efficient. Sillem discovered that better time resolution was achievable with high frequency elements than for Fourier methods but the results were not dramatically better in terms of frequency resolution.

The next stage in his study was to extract a set of features that represent continuous sounds with sharply defined frequencies. He constructed the features by connecting a frequency peak in one time-slice with a similar or identical one in the next time-slice. At the end of this step a large list of raw frequency features existed, however not all were reliable in the current state of the study.

Future developments included tests to see if instruments could be recognised by comparing the amplitude and phase of the note harmonics to those of actual instruments. Once the notes, onset and offset times and instrument type have been identified, the data would then be output into a MIDI file. However, this project had not been completed, even as a feasibility study.

[Chan '00] described a system that gives real time automated transcription of live monophonic music into sheet music using common music notation. They attempted to implement a system on the TMS320C67x EVM board to accept live music through a microphone, detect the notes and transcribe them into sheet music in real time. The constraints of the project were that the range of detection was limited to two octaves (C4 – C6), a note could be no shorter than a semi crochet and the input signal must not have a high noise content.

They decided to use a simple time-domain approach as opposed to a frequency-domain technique to find the pitch period of the musical signal. [Wendt '96] determined the Glottal Closure Instants in speech signals and measured the time period between each event. A Glottal Closure Instant is a vocal string effect occurring during each period of voiced speech when the glottis is excited. The signal was processed period by period which provided impressive time and frequency resolution. The method proposed was to use a single derivative filtering function defined to contain a specific bandwidth of voiced speech. When this wavelet function was convolved with the voiced speech signal, a filtered signal with well-defined local maxima where Glottal Closure Instants occur was produced. This method promised dramatic simplification over methods previously used on speech signals in terms of processing as only convolution was used.

A Glottal Closure Instant is equivalent to a zero crossing in a speech signal. Therefore, [Jehan '97] investigated the adaptation of this algorithm towards musical signals. By filtering a musical signal with a derivative function a maximum will occur at each zero crossing with the time period between each representing the pitch period of the signal at that moment. It was this algorithm that was implemented by Chan. The wavelets and the final derivative function were chosen by utilising a Daubechies wavelets [Jense '01] generating function and investigating the properties of the resulting wavelet. In general, the higher the order of the filter, the longer it is and the better it performs. However, increasing the order also increases computational expense, so an eighth order filter was chosen and a Matlab function was used to find the coefficients of the Daubechies wavelets. The coefficients were passed into another function that generated the wavelet and scaling functions, the frequency band was also an input. However, Chan discovered that the wavelets alone were not satisfactory

as band-pass filters since the side-lobes were of relatively high magnitude and therefore allowed some of the next harmonic to pass through.

It was then decided to use a combination of wavelets and a Finite Impulse Response (FIR) filter in order to improve the problem. Low pass FIR filters were convolved with the wavelets to suppress the side-lobes further. They were designed so that the cut-off frequency was in between the last note of the octave of interest and the first note of the next octave. These filters were substituted for the ones used in [Jehan '97] and the results of the pitch detection became extremely reliable.

Chan used the parabolic approximation technique that was used to locate the maxima in [Jehan '97], since simply choosing the visible peaks may not locate the actual maxima due to sampling restrictions. The parabola fits three points and from these the maximum was located, this is similar to the peak interpolation method used in sinusoidal modelling. There were some unwanted local maxima due to noise, which were eliminated by checking the variation of two consecutive durations and the dynamic threshold. The samples that did not fit into a specified range were eliminated provided there was not a really strong change in the pitch. To reject un-pitched notes and errors, they calculated the energy of each 512-sample frame of the input and the filtered signal (using the sum of squares method). Rejection occurred when the input signal energy was below a set threshold, if there was no energy in the filtered signal or if the pitch track exhibited rapid variations. Harmonics are present in the majority of all pitched musical sounds, i.e. when a note is played its harmonics in the higher octaves can also be heard. In this case, they solved the problem by always selecting the lowest note detected.

The Matlab code developed by Chan was converted to C code. This project was dependent on the user supplying the tempo, number of beats per minute, shortest note duration and the clef. When all of the musical elements were identified – pitch, duration, meter – this information was transcribed into musical notation using Common Music Notation [Schottstaedt].

[McNab '00] gives an evaluation of the melody transcription system used in MelDex (Melody Index), a system that was developed in the University of Waikato, Hamilton, New Zealand. MelDex enables users to search for musical scores within a collection

by singing short phrases into a microphone. The users must sing the melody using the syllable "da"; this melody is recorded and then transcribed into music notation.

The signal was segmented into the individual notes using the Root Mean Square (RMS) power of the signal. Two thresholds were used to accommodate noise. These are calculated from the second-order RMS power over the entire input sound. When the amplitude exceeds the higher threshold, a note onset is detected and when it drops below the lower threshold, a note offset is detected. If the note segment is found to be less than a third of the shortest allowed, set by the user, it is discarded. The frequency is determined using the Gold-Rabiner algorithm, which is described in [Gold '69]. This algorithm gives a frequency estimate for each pitch period and McNab's implementation averages these over 20ms frames. The frequency of each note segment was determined using a histogram.

Ten people, five male and five female, were given a set of eleven well-known songs to sing in order to test the system. One of the subjects did not know one of the songs so that song was dropped from the test. Therefore the implementation was finally tested on a set of 100 examples. The subjects had varying levels of singing experience ranging from negligible to extensive. The user was able to visually inspect the segmentation points and manually reposition them and listen to the segments if necessary. Segmentation errors were recorded and these fell into four categories, insertion, deletion, concatenation and truncation. There was an error of 11.3%, mainly concatenation errors, over half of which occurred on short notes. A big problem was finding onset and offset threshold values that covered all situations. Frequency errors were classed as the wrong octave and the wrong note and had an error rate of less than 1%. These are very accurate results but this was due to the use of a histogram, which finds the average frequency during a note segment. Frequency results at the frame level did contain errors. This means that frequency determination relies on the note segmentation information in order to give accurate results, however the note segmentation results were not accurate enough. Therefore, the conclusion resulting from the evaluation was that it would be necessary to replace the Gold-Rabiner algorithm in order to obtain better frequency results at the frame level and to enable note segmentation at that stage.

[Monti '00] describes a transcription system that uses an autocorrelation based pitch

tracker. They listed four physical parameters that can be used to describe a musical signal, along with how a listener perceives these parameters:

1. Fundamental frequency or pitch.
2. Signal amplitude or loudness.
3. Signal shape or timbre.
4. Sound source location or spatial perception.

The fundamental frequency of each note was calculated using the autocorrelation function in Equation (3.5). This function shows peaks for any periodicity present in the signal and the zero lag autocorrelation gives the energy of the signal. Some algorithms from [Slaney '98] were used in the implementation. There was good detection for the steady part of the note since the pitch remains almost constant while useful information could not be provided during the noisy attack transient. The envelope of the signal was calculated to assist the pitch tracker and the pitch calculation was skipped whenever the energy of the signal dropped below the audibility threshold. Each pitch was then converted to its corresponding MIDI number.

The MIDI numbers and signal envelope were then passed into a collector, which extracted the score. The main parameter of the collector was the minimum note duration. When a pitch maintained the same value for the minimum note duration, a note onset was detected. This was modified to adapt to the speed of the music, which improved the performance of the system. A note offset was detected by a change in frequency or silence. If the pitch varied after a note onset before returning to the original pitch within the minimum note duration, it was considered to be one note. Note duration was calculated as the difference between its onset and offset times.

[Csound] was used to synthesize the transcribed music using the pitch, onset and duration of the notes, which were obtained from the collector. In addition to these three parameters, the envelope and timbre were also required to recreate the original melody. Each note envelope was recreated using the Csound function *linseg*, which traces linear segments between defined points. This was useful in detecting missed notes as two notes of the same frequency played consecutively would be detected as one by the collector. When the amplitude rose twice for the one note, it was assumed

that a second note of the same frequency had been played. The timbre of a note depends on the characteristics of the instrument. To recreate the timbre, the basic waveform of each note was extracted from the original signal and multiplied by its envelope. The section of waveform was chosen after the initial attack had occurred, when the note was in its steady state. However since the note transients were not considered, poor synthesis results were achieved.

The system was tested on a CD collection of brass instrument riffs and the tempo and pitch were correctly extracted. The results were compared with those of a commercial program [WAV2MIDI] with similar results being achieved.

[Plumbley '02] found that in practice using the Fast Fourier Transform (FFT) [Cooley '65], was more efficient than the autocorrelation function. The autocorrelation measures the similarity between shifted versions of the time waveform and the delay corresponding to the highest peak in the autocorrelation gives the period of the waveform. Since frequency is the inverse of time period, this information can be used to give the frequency of the signal. Since high frequencies have a short duration in the time domain, Plumbley found that using a frequency transform gave more accurate results.

[Bello '00] gave an overview of two techniques for automatic music transcription, which were developed at Kings College in London. The first was a monophonic transcription system, described above in [Monti '00], which was tested on a saxophone and a flugel horn. He found that it had good detection and smooth values during the steady part of the note, i.e., after the initial attack of the note onset, and gave correct results for the note range it was tested on, B1-E6.

The second of their techniques was a simple polyphonic transcription blackboard system and Neural network. Top-down processing, where different levels of the system are determined by predictive models of the analysed object or by previous knowledge of the nature of the data, was used in this project and was achieved through the implementation of a neural network [Haykin '99]. It was noticed that while analysing the running spectrum of the sound, when an onset occurs there is a large burst of energy, particularly in the high frequencies. They made use of this

property to determine onset time. Segmentation was performed by averaging the STFT of the signal between onsets.

The levels in this blackboard system were:

1. Tracks, the peaks of the averaged STFT in a given segment.
2. Partials, which were created from the tracks and created a link between tracks and notes.
3. Notes representing the high-level musical structures the system aims to extract.

A feed forward network was the type of neural net implemented. The input pattern consisted of the spectrogram of a piano signals segment (a note/chord). The target output was represented by the absence "0" or presence "1" of a chord in the sample.

When the system was running, the network received the STFT data, which the blackboard system analysed as input. The networks output changed the performance of the system allowing multiple note hypotheses to survive if necessary. The output was in the form of a piano roll and *CSOUND* score file. A recurring problem was that the system chose a note an octave below the actual note that was played. It was suggested that new knowledge sources should be added to overcome this octave detection problem. The architecture needed to be modified, to incorporate dynamic structures to handle different sized hypotheses, i.e. chords with more than three notes. The training space also needed to be expanded to include all eight octaves of the piano.

[Goto '00] proposed a system, called PreFEst (Predominant F0 Estimation Method), to estimate the fundamental frequency of melody and bass lines in CD recordings. The system did not rely on the presence of the fundamental frequency in order to calculate the pitch of a note and instead obtained the most predominant fundamental frequency, which was supported by predominant harmonics within a specified frequency range. PreFEst made the following assumptions:

- The melody and bass sounds contain the harmonic structure of a tune, however did not care about the presence of the F0's frequency component.

- The melody line is predominant in the middle and high frequency regions and the bass line predominant in the low frequency region.
- It was also assumed that the melody and bass lines tend to have temporally continuous trajectories.

PreFEst first calculates instantaneous frequencies using multirate signal processing techniques and extracts candidate frequency components on the basis of an instantaneous-frequency-related measure. PreFEst basically estimates the F0, which is supported by predominant frequency components within a limited frequency range. The frequency range was limited by using two band-pass filters, one for the middle and high regions of the melody line and a second for the low region of the bass line. It then forms a probability density function of the F0 and this represents the relative dominance of every possible harmonic structure.

The time-frequency representation of the signal was obtained using an STFT-based multirate filter bank. Initially the signal was down-sampled using a decimator, which contained an FIR low pass filter with a cut-off frequency of *0.45fs*, where *fs* was the sample rate of the branch. This resulted in five output branches – from the original signal at 16 kHz to the final decimated signal at 1 kHz – with each of these branches analyzed as follows. Firstly the STFT of a branch was obtained:

$$X(\omega,t) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-j\omega\tau}d\tau = a + jb \qquad (3.8)$$

where *x(t)* is the time domain signal and *h(t)* is the window function. Time delays at the different multirate layers were compensated for. Potential frequency components were extracted based on mapping from the center frequency, ω, of a STFT filter to the instantaneous frequency, $\lambda(\omega,t)$, of its output [Flanagan '66]:

$$\lambda(\omega,t) = \omega + \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2 + b^2} \qquad (3.9)$$

Once fixed stable mapping points were found, it was then possible to extract a set of instantaneous frequencies of the frequency components as follows [Abe '97]:

$$\Psi_f^{(t)} = \{\psi \mid \lambda(\psi,t) - \psi = 0, \frac{\partial}{\partial \psi}(\lambda(\psi,t) - \psi) < 0\} \quad (3.10)$$

Calculating their powers, the power distribution function was defined as:

$$\Psi_p^{(t)}(\omega) = \begin{Bmatrix} |X(\omega,t)| if \omega \in \Psi_f^{(t)} \\ 0 \qquad otherwise \end{Bmatrix} \quad (3.11)$$

The frequency range was limited using two band pass filters. One covered a low frequency range, 1000 - 4800 cents, to correspond with dominant bass line harmonics and the second, a middle to high frequency range, 3600 – 9600 cents, to correspond with dominant melody line harmonics. There are 100 cents to a semi-tone and frequency in cents, f_{cent}, is obtained from frequency in Hz, f_{Hz}, as follows:

$$f_{cent} = 1200 \log_2 \frac{f_{Hz}}{440 \times 2^{\frac{3}{12}-5}} \quad (3.12)$$

The filtered frequency components could then be represented as $BPF_i(x)\Psi_p'^{(t)}(x)$, where $BPF_i(x)$ was the frequency response of the band pass filter at x (in cents) for the melody ($i=m$) and bass ($i=b$) lines. The power distribution was the same as in Equation (3.11) above with the exception that frequency was now in cents. The probability distribution function (pdf) of each of the filtered frequency components was given by:

$$p_\psi^{(t)}(x) = \frac{BPF_i(x)\Psi_p'^{(t)}(x)}{\int_{-\infty}^{\infty} BPF_i(x)\Psi_p'^{(t)}(x)dx} \quad (3.13)$$

giving a pdf of the fundamental frequency. It could be considered that the pdf was created from a weighted mixture of harmonic-structure tone models. The pdf of each tone model was denoted as $p(x|F)$ where F was its fundamental frequency and the mixture density was defined as:

$$p(x;\theta^{(t)}) = \int_{Fl_i}^{Fh_i} w^{(t)}(F)p(x \mid F)dF, \quad (3.14)$$

$$\theta^{(t)} = \{w^{(t)}(F) \mid Fl_i \le F \le Fh_i\} \tag{3.15}$$

where Fl_i and Fh_i were the upper and lower limits of the possible fundamental frequency range, and $w^{(t)}(F)$ was the weight of a tone model that satisfies $\int_{Fl_i}^{Fh_i} w^{(t)}(F)dF = 1$. All simultaneous possibilities of the fundamental frequency needed to be considered, as it was not known how many sound sources were contained in the mixture. By estimating $\theta^{(t)}$ so that $p_v^{(t)}(x)$ was likely to have been created from $p(x;\theta^{(t)})$, $w^{(t)}(F)$ could be interpreted as the pdf of the fundamental frequency.

To estimate the value of $\theta^{(t)}$, the Expectation-Maximization (EM) algorithm described in [Dempster '77] was used. This algorithm consists of two iterative steps and is used to calculate maximum likelihood estimates from incomplete data. The first step is the Expectation or E-step and calculates the conditional expectation of the mean log-likelihood as follows:

$$\Theta(\theta^{(t)} \mid \theta'^{(t)}) = \int_{-\infty}^{\infty} p_v^{(t)}(x) E_F[\log p(x, F; \theta^{(t)}) \mid x; \theta'^{(t)}]dx \tag{3.16}$$

where $E_F[a|b]$ is the conditional expectation of a with respect to the hidden variable F with the probability distribution determined by the condition b. The second step is the Maximization or M-step, which maximizes $\Theta(\theta^{(t)} \mid \theta'^{(t)})$ as follows:

$$\overline{\theta^{(t)}} = \arg\max_{\theta^{(t)}} \Theta(\theta^{(t)} \mid \theta'^{(t)}) \tag{3.17}$$

to obtain the 'new' estimate $\overline{\theta^{(t)}}$ from the 'old' estimate, $\theta'^{(t)}$, which was updated on each iteration.

The peak trajectories, resulting from the EM algorithm, were sequentially tracked to select those which were the most dominant and stable. This was achieved using a multiple-agent architecture. A salience detector extracted salient peaks from the pdf of the fundamental frequency. Agents were generated which interacted and allocated peaks amongst themselves. When more than one agent claimed the same peak, it was

allocated to the most reliable agent. When the most salient peak was not located, a new agent was generated to track it. If an agent had not been allocated a salient peak, or could not find its next peak in the pdf it was penalised and once it was allocated a peak, the penalty was reset. Once this penalty exceeded a threshold, the associated agent was terminated. Each agent evaluated its own reliability using the reliability at the previous frame and the degree of the peaks salience at the current frame. The most reliable agent, and the power along the trajectory of the peak it was tracking determined the final value of fundamental frequency. The system was tested on excerpts of ten songs with an average accuracy of 86.5% achieved for melody detection and 75.3% for bass detection.

[Klapuri '00] proposed a multipitch estimation (MPE) algorithm to operate reliably on polyphonic music. The algorithm was divided into two parts, the first involved estimating the predominant pitch and the second subtracting the spectrum of this detected pitch from the mixture.

Firstly a frequency representation of the signal was obtained by applying a Discrete Fourier Transform (DFT) to the Hamming-windowed signal. An enhanced spectrum, $X_e(k)$, was calculated by taking the logarithm of the magnitude spectrum and high pass filtering the result. The signal was then separated into 18 logarithmically distributed bands. Each band covered two thirds of an octave, weighted by a triangular window with 50% overlap between bands. A fundamental frequency likelihood vector, $L_B(n)$, was calculated for each band, B, so that each frequency sample of $X_e(k)$ had a corresponding fundamental frequency likelihood sample of $L_B(n)$. Frequency samples $X_e(k)$ at band B were in the range $k\ [k_B,\ k_B+K_B-1]$ where k_B is the lowest sample and K_B the number of samples at the band. The bandwise fundamental frequency likelihoods, $L_B(n)$, were calculated by finding such a series of every n^{th} spectrum sample that minimises the likelihood:

$$L_B(n) = \max_{m \in M} W(H) \sum_{h=0}^{H-1} X_e\left(k_B + m + hn\right) \qquad \textbf{(3.18)}$$

where $m \in M,\ M=\{0,1,\ldots,n-1\}$ is the offset of the series of partials, $H = \lceil (K_B - m)/n \rceil$ is the number of harmonic partials in the sum and $W(H)=0.75/H+0.25$ was used as a normalization factor.

Before the bands could be recombined, the inharmonicity factor was calculated. Inharmonicity is explained in chapter 1. This was because the fundamental frequencies at different bands did not match for inharmonic sounds. Inharmonicity caused a slight rise in the perceived pitch. It was also discovered that raising the likelihoods to a second power, prior to recombining, gave better robustness. The outputs of this stage were the fundamental frequency, inharmonicity factor and the frequencies and amplitudes of the harmonic series of the sound.

In the second stage of the system, the partials of the detected sounds were removed from the mixture. Although good estimates of the amplitude, a_s, angular frequency, ω_s and phase, θ_s, of each partial, $s(t) = a_s \cos(\omega_s t + \theta_s)$, of a sound were found in the first section, more accurate estimates were found by applying a Hamming window and zero padding in the time domain and then using quadratic interpolation of the spectrum. The STFT of each partial, $s(t)$, was obtained as follows:

$$S(\omega) = \int_0^T \left[w(t)s(t)e^{-it\omega} \right] dt \tag{3.19}$$

where temporal weighting was performed by the window function:

$$w(t) = a_w + b_w \cos(\omega_w t + \theta_w), t \in [0, T] \tag{3.20}$$

This partial was then subtracted from the mixture. This process was repeated for each partial of the detected sound.

A problem was encountered when sounds with overlapping partials were contained in the same sample of the mixture. This resulted in a corrupted mixture when partials of sounds, which had not yet been subtracted, had already been removed. To rectify this, the sound spectrum was smoothed, by calculating a moving average over the amplitudes of the harmonic partials. A hamming window, an octave wide, was centred at each harmonic and a weighted mean calculated within this window. Then the minimum amplitude of the original and average was taken as the new amplitude. This meant that the partials of the sound would have much lower amplitudes, and would not be completely removed from the mixture. This adjustment approximately halved error rates in polyphonies.

The system was tested on a database of sung vowels and twenty-six different musical instruments, which included plucked and bowed strings, flutes, brass and reed instruments. Sounds were played in a five-octave range from 65 Hz to 2100 Hz and pitches were estimated in a single 190ms time frame. Results were compared with those obtained manually by 10 musicians. Overall, the implemented system performed better than the trained musicians.

[Klapuri '01] combined existing algorithms, the onset detector previously presented in [Klapuri '99] (see section) and the multipitch estimator in [Klapuri '00], and added two new features to enable the application to deal with real world signals. The features were both part of the MPE algorithm, the first was noise suppression, which occurs before the predominant pitch estimation, and the second estimated the number of voices present in the sound mixture.

In a musical signal, noise is typically caused by drums and percussive instruments. Since this is not continuous throughout the entire signal, the proposed system estimated and removed noise independently in each analysis frame. The system dealt with both convolutive and additive noise which are defined as follows:

$$X(k) = S(k)H(k) + N(k) \tag{3.21}$$

where $X(k)$ is the power spectrum of the input signal, $S(k)$ the power spectrum of the vibrating system whose fundamental frequency is to be measured, $H(k)$ is the frequency response of the operating environment and body of the instrument and filters $S(k)$, and $N(k)$ is the power spectrum of the additive noise. Both additive and convoluted noises were removed simultaneously, using a technique based on RASTA spectral processing in [Hermansky '93]. The power spectrum was transformed as follows:

$$Y(k) = \ln\{1 + J \times X(k)\} \tag{3.22}$$

where J scales the input spectrum so that $N(k) << 1$ and the spectral peaks of $[S(k_p)H(k_p)] >> 1$ (k_p corresponded to the fundamental frequency of spectral peak p). J depended on the level of additive noise and the spectral peaks and was calculated as follows:

$$J = \alpha \left(\frac{k_1 - k_o}{\sum_{k=k_0}^{k_1} X(k)^{1/3}} \right)^3 \tag{3.23}$$

where k_0 and k_1 are determined by the frequency range, in this case they were 50 Hz and 6 kHz, and the optimal value for α was 1. *M(k)*, the noise content of *Y(k)*, was estimated by calculating a moving average over *Y(k)* in equivalent rectangular bandwidth (ERB) frequency scale [Smith '04]. This was achieved by calculating a Hamming-window weighted average over *Y(k)* for values around k where the width of the window was dependent on the centre frequency f corresponding to k and was calculated as follows:

$$W(f) = \beta \times 24.7 \left(4.37 \frac{f}{1000} + 1 \right) \tag{3.24}$$

where the optimum value for β was 4.8. This estimated noise spectrum, *M(k)*, was linearly subtracted from *Y(k)* resulting in *Z(k)* with resulting negative values set to zero. This value was then passed on to the multipitch estimator.

In a study [Huron '89] on the ability of musicians to identify the number of voices present in a polyphonic mixture, there was a noticeable drop in accuracy when there were four or more voices present, with the musicians tending to underestimate the number of voices in over half of the examples. Klapuri used a statistical approach to solve the problem. Random mixtures were created that contained between zero and six simultaneous harmonic sounds from 26 musical instruments, and these were contaminated with pink noise or random drum sounds. These mixtures of known polyphony were used to test the system, with different characteristics being measured in order to find a way of determining when to stop the iterations. It was determined that two techniques were necessary to estimate polyphony; the first detected if there were harmonic sounds in the signal and the second how many simultaneous sounds were present, if any.

Voicing was detected using the likelihood, L_1, calculated in the first iteration of the predominant pitch estimator, using Equation (3.18). This was combined with features related to the signal-to-noise ratio of the input signal:

$$V_0 = 4\ln(L_1) + \ln\left(\frac{P_x}{P_M}\right) \tag{3.25}$$

where P_x is the power spectrum of $X(k)$ between 50 Hz and 6kHz after the signal had been scaled with J and P_m is the power spectrum of the estimated noise spectrum in the same frequency region, calculated by transforming $M(k)$ back to power spectral domain by inverse transforming Equation (3.17). The signal was voiced if $V_0 > T_{voicing}$, where $T_{voicing}$ was a fixed threshold.

To stop the iterative multipitch estimator, the likelihoods, L_i, of the predominant pitch estimator were again used. The likelihoods are affected by the signal-to-noise ratio, decreasing as noise increases. This is corrected using the following equation:

$$V_i = 1.8\ln(L_i) - \ln\left(\frac{P_x}{P_M}\right) \tag{3.26}$$

when V_i remains greater than a fixed threshold, the iterative process continues.

Results for both voicing and the number of iterations in the multipitch estimation system can be calculated without using the likelihoods of the predominant pitch estimator as follows:

$$V_0' = \ln\left(\frac{P_x P_z}{P_M}\right) \tag{3.27}$$

$$V_i' = 2\ln(P_i) - \ln\left(\frac{P_x}{P_M}\right) \tag{3.28}$$

where P_z in Equation (3.27) is the power of $Z(k)$, and P_i in Equation (3.28) is the power of the sound detected at iteration i, calculated by selecting frequency samples from $Z(k)$ from the positions of the harmonic components of the detected sound, transforming to the power spectral domain and summing. The accuracy of this method was comparable to using the likelihoods.

The biggest problem for the voicing detector was the presence of drum sounds. As half of the acoustic energy of the sounds of bass drums, snares and tom-toms is

harmonic, the voicing detector tended to be misled. This system was tested on simulated examples as opposed to real musical signals, which are significantly more complex. It was thought that MIDI-songs would give the best evaluation of the system and initial tests had not yet been completed.

[Marchand '01] proposed a pitch tracker that used a series of Fourier transforms. To determine the pitch at a time, t, this section of the signal was multiplied by a Hanning window and then analysed using the order-1 Fourier transform. The fundamental frequency of a note may be determined from this analysis using the equation:

$$f = \frac{F_s i_{FT}}{N} \tag{3.29}$$

where f is the fundamental frequency, i_{FT} is the bin number of the peak of greatest magnitude in the Fourier transform, F_s is the sampling frequency and N is the size of the Fourier transform. The problem with using this method is that occasionally, the fundamental frequency is missing so the peak of greatest magnitude may correspond to a harmonic giving the wrong result.

Therefore, the magnitude spectrum of the first Fourier transform was analysed using the classic Fourier transform, resulting in a Fourier of Fourier transform of the original signal. This time, the fundamental frequency is calculated as follows:

$$f = \frac{F_s}{2i_{FT(FT)}} \tag{3.30}$$

where $i_{FT(FT)}$ is the bin number of the peak of greatest magnitude (after that at bin 0) in the second Fourier transform. However, this set of peaks did not always give the correct frequency value so a peak-tracking strategy similar to that used in [Althoff '99] was applied to deal with the pseudo-partials detected in the Fourier of Fourier transform spectrum. Partials are eliminated according to amplitude and length with the dominant ones remaining. The frequency of the dominant partial is then taken to be the fundamental frequency.

The system was tested on many natural sounds like the saxophone, guitar and human voice with good results being achieved. The algorithm was found to be much faster than the autocorrelation method. The results were compared to those obtained using the system described in [Arfib '99]. Error rates of 1-6% were found for the proposed system compared with error rates of 5-12% for the [Arfib '99] system.

This review on music transcription systems has described some of the current techniques that are being used to tackle the problem of automatic music transcription. Time-domain domain techniques, such as autocorrelation, are fine when dealing with monophonic transcription as they deal with the audio signal as a whole but are not suitable for polyphonic music. Frequency-domain techniques, such as the STFT, have proved to be more popular, are more computationally efficient and are also more effective.

3.2 Onset Detection

It became clear during the literature review that a good onset detector is an essential element of a music transcription system. An onset indicates when something new has happened in a signal. In the case of a musical signal, this could be a change in pitch, i.e. the precise time when a new note is produced by an instrument. Onsets are very important in instrument recognition, as the timbre of a note with the onset removed has proved be very difficult to recognize [Sethares '97]. Knowing the location of the onsets allows a signal to be segmented so that each event can then be examined in isolation. Therefore, in order to determine what notes constitute a musical piece, an onset detector is vital. Onset detectors are not exclusive to music transcription as they are also used in a number of other applications such as music instrument separators, time stretching and instrument recognition as mentioned above.

There are different types of onset, the most common of which can be loosely termed fast and slow onsets. A fast onset is the easier to recognize since it has an abrupt change in the energy profile for a short duration at the beginning of a note and the onset is very noticeable in the higher frequencies. This type is typical of the percussive instruments such as a piano or a banjo. An example of a fast onset, played by a piano, is shown below in figure 3.5 (a).

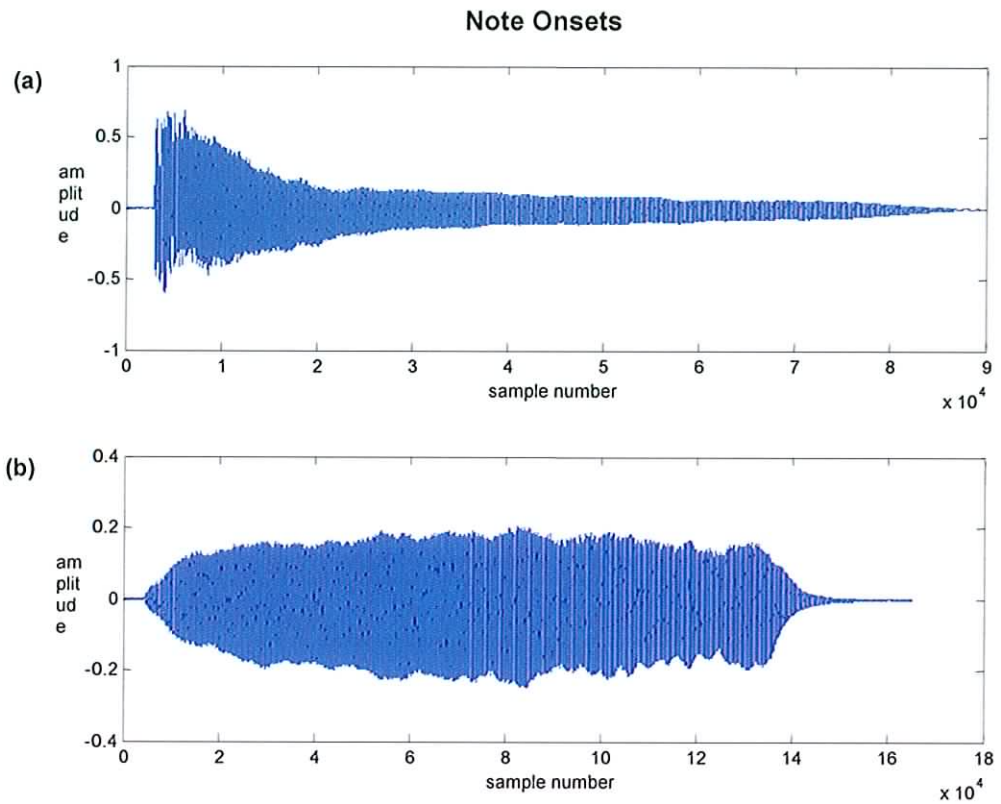

Figure 3.5: (a) A single note played by a piano; (b) A single note played by a fiddle

The slow onset is much harder to recognize as it takes a much longer time to reach the maximum onset value and has no noticeable change in the higher frequencies. This type is typical in wind instruments such as the flute and tin whistle and bowed instruments such as the fiddle. An example of a slow onset, played by a fiddle, is shown in figure 3.5 (b).

A great deal of research has been carried out in the area of onset detection, and although much progress has been made, it still remains an unsolved problem, particularly in the case of slow onsets. The first onset detectors considered the musical signal as a whole [Chafe '85]. However this approach is only suitable for monophonic signals with very prominent onsets. [Bilmes '93] introduced a sub-band approach by dividing the signal into two bands. He calculated the short time energy of the high band using a window and then computed the slopes of the energy over time, searching for a value that reached a given threshold. Once the threshold had been attained, the attack time was the maximum slope in a predefined region.

[Masri '96] proposed an approach to detect onsets in order to improve music analysis-resynthesis. He based his approach on the deterministic plus stochastic sinusoidal model developed by [Serra '90]. The energy, E, and high frequency content, HFC, of the k^{th} bin of the Fourier transform of the signal, $X(k)$, were measured as follows:

$$E = \sum_{k=2}^{N/2} \left\{ |X(k)|^2 \right\} \tag{3.31}$$

$$HFC = \sum_{k=2}^{N/2} \left\{ |X(k)|^2 . k \right\} \tag{3.32}$$

where N is the FFT array length. The condition for onset detection was as follows:

$$\frac{HFC_r}{HFC_{r-1}} * \frac{HFC_r}{E_r} > T_D \tag{3.33}$$

Where r is the current frame and T_D the threshold above which an onset is detected.

[Scheirer '98] built upon the sub-band approach, which is more suitable for dealing with polyphonic signals. He carried out psychoacoustic analysis on musical signals to better understand beat perception. A musical signal was divided into six frequency bands and their energy envelopes calculated. These were modulated with the corresponding frequency bands of a noise signal and the resulting bands summed together. It was discovered that the resulting noise signal possessed a rhythmic percept that was significantly the same as that of the original musical signal. This implies that only the energy envelopes of a musical signal are required to extract pulse and meter information. However, it is necessary to divide the signal into bands and process each one individually, otherwise the rhythmic content is not preserved. Based on this information, he divided a musical signal into six bands with the range of each band roughly covering one octave. This was achieved using sixth order band-pass elliptic filters with the lowest being a low-pass filter and the highest, a high pass filter. The amplitude envelope was extracted from each band, and this was convolved with a 200ms half-hanning window in order to further smooth the signal. Convolving with a half-hanning window performs much the same task as the human auditory system, masking fast amplitude modulations but emphasising the most recent inputs. The first

order difference function was calculated and then half wave rectified. Results for extracting the beats of musical signals were similar to those obtained by human listeners. However, the system was only suitable for use on percussive sounds, i.e. sounds with fast onsets.

[Klapuri '98] used Schierer's approach as a basis for his own system, which detects the onset of notes in a musical signal. This time the musical signal was divided into seven one-octave bands. The first order difference function of each of the amplitude envelopes was calculated and only the segments that were above a certain threshold were considered. The first order relative difference function, *W(t)*, was then calculated:

$$W(t) = \frac{\frac{d}{dt}(A(t))}{A(t)} = \frac{d}{dt}(\log(A(t))) \qquad (3.34)$$

where *A(t)* is the amplitude envelope of the signal. This measures the amount of change in relation to signal level and is the same as differentiating the logarithm of the amplitude envelope. It gives a more accurate onset time than using the first order difference function. This is illustrated below in figure 3.6.

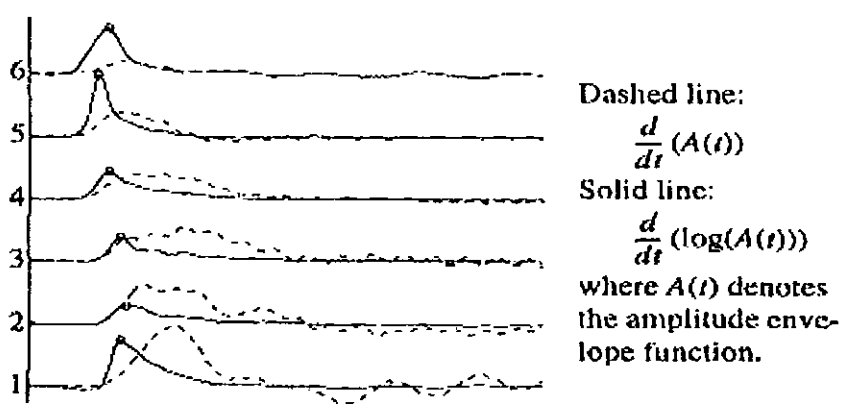

Figure 3.6: First order *absolute* (dashed) and *relative* (solid) difference functions of the amplitude envelopes of different frequency bands. Picked onset times are circled [Klapuri '98]

Potential onset times were then collected from each band by performing a peak picking operation on the relative difference function. Finally, the results from the different frequency bands were combined to give the onset times of the signal.

[Klapuri '99] describes a modified version of the onset detector described in [Klapuri '98]. This time he incorporated the psychoacoustic loudness perception model described in [Moore '97]. A bank of twenty-one nearly critical-band filters were used in an attempt to mimic the human auditory system. The filters covered the frequency range from 44 Hz to 17 kHz, the lowest three were one-octave band-pass filters and the remaining eighteen were third-octave band-pass filters. As in the previous case, the calculations were performed one band at a time. The outputs of the filters were full wave rectified and decimated by a factor of 180 in order to ease the following computations. This time the signals were convolved with 100ms half Hanning windows in order to calculate the amplitude envelopes. The first order relative difference function was calculated in the same way as in the previous system. The onsets were then detected by using a simple peak picking operation, which looks for peaks above a global threshold.

The intensity of each onset was extracted from the first order difference function. It was chosen between the onset and the point where the amplitude envelope begins to decrease, at the point of maximum slope. Any component that was detected less than 50ms from a more intense component was discarded. The onset components in the different frequency bands were then sorted in time order. Each onset was then assigned a loudness value, which was calculated by collecting onset intensities in a 50ms time window around the onset components. The loudness results compared very well with the perceived loudness of onsets in listening tests. A loudness threshold was then set and any components with a value below this were dropped. Components that were less than 50ms from a 'louder' component were also dropped.

The system was tested on musical signals containing a range of instruments and genres. The period of each signal was inspected and their onsets marked. The signals were then fed into the system and those results compared with the manual ones. Good results were achieved for signals containing a small number of instruments but it performed very poorly on symphony orchestras, as the system was not able to deal well with strong amplitude modulation.

[Marolt '02a] and [Marolt '02b] presented a system to detect note onsets in polyphonic piano music. The onset detector was a part of their transcription system, SONIC. It was based on a model for segmentation of speech signals in [Smith '96], which uses a network of integrate and fire neurons [Gerstner '02]. Marolt added a multilayer perceptron (MLP) neural network to improve reliability.

The signal was split into 22 frequency bands using band-pass infinite impulse response (IIR) filters. These filters were designed to imitate the function of the basilar membrane in the human inner ear. The signal in each of the frequency bands was processed by a filter, which calculated the difference between two amplitude envelopes:

$$O(t) = \int_0^t \left(\exp\left(-\frac{t-x}{f_s t_s} \right) - \exp\left(-\frac{t-x}{f_s t_l} \right) \right) s(x) dx \qquad (3.35)$$

where $s(x)$ is the signal, f_s the sampling frequency and t_s and t_l the time constants of two smoothing filters. t_s is a short time constant of 6-20ms, depending on the centre frequency of the signal, and t_l a longer time constant of 20-40ms. The output of the difference filter is positive when the signal energy rises and negative otherwise.

Peak picking was performed using a combination of a network of integrate and fire neurons and an MLP. Neural networks are modelled on how biological nervous systems, such as the brain, process information. Most biological neurons communicate by short electrical pulses. In artificial networks, integrate and fire neuron models do not rely on a temporal average over the pulses, unlike the standard neuron model. Each neuron, i, changed its activity, A_i, which was initially set to zero, as follows:

$$\frac{d}{dt} A_i = O_i(t) - \gamma A_i \qquad (3.36)$$

where $O_i(t)$ is the output of the i^{th} difference filter and γ describes the leakiness of integration. Once A_i reached a threshold, a neuron was fired, i.e. emited an output pulse, and A_i was reset to zero. The firing of neurons was an indication of amplitude growth in each band. Once a neuron had been fired, the next input could not be

accepted for a refractory period of 50ms. All neurons were connected via excitatory connections and once one was fired, activities of all others were raised.

The MLP neural network was used to determine which of the pulses are note onsets. Its inputs were the neuron activities, A_i, and several other factors such as the amplitudes of the frequency bands. The system was trained to recognise note onsets on synthesised piano tunes and tested on both real and synthesised polyphonic piano tunes with good results being achieved. They found that network connections improved detection of weak onsets as they encourage neurons close to the threshold to fire. The refractory period meant that a series of spurious onsets were not detected for the one note, however, it also meant ornamentation and some notes played in quick succession went undetected.

[Duxbury '02] presented an approach to onset detection that used an energy-based detector on the upper sub-bands and a frequency-based distance measure on the lower sub-bands. A method for improving the detection function by using a smoothed difference metric was also presented and they showed that by analysing the statistics of this detection function, the detection threshold could be set automatically.

Using a constant-Q filter bank, the signal was split into five bands. The highest band was disregarded as it contains weak onset information. The next three bands (1.2-11kHz) contain bursts of energy for a range of onsets while although the lowest band (0-1.1kHz) does not, it does have noticeable differences in the frequency content at note changes. An energy-based technique was used on the upper three bands (1.2 – 11KHz). The sub-band energy, *SE*, was given by:

$$SE(n) = \sum_{m=(n-1)h}^{nh} |x(m)|^2 \qquad (3.37)$$

where m is the time index, n the hop number and h the hop size.

A transient energy measure was also used in the upper bands to eliminate upper steady state components such as high frequency partials or high-pitched notes. To do this, an extra STFT was used in each band. This was based on basic phase vocoder theory [Dolson '86]. It is possible to show that when presented with a perfect sinusoid

in ideal conditions, the instantaneous phase should be equal between successive frames of a STFT. However this is an unrealistic assumption in real conditions and the unwrapped estimate phase, $\tilde{\varphi}$, of the k^{th} bin is expected to be equal to the target phase, $\tilde{\varphi}_t$, plus a deviation phase, $\tilde{\varphi}_d$, i.e.:

$$\tilde{\varphi}_d(m,k) = princarg\left[\tilde{\varphi}(m,k) - \tilde{\varphi}_t(m,k)\right] \tag{3.38}$$

Where *princarg* is the principal argument function mapping the phase to the [-π,π] range and m is the hop number. The instantaneous frequency, f_i, of the k^{th} sinusoid is calculated by dividing the unwrapped phase difference, $\Delta\tilde{\varphi}(m,k)$, between consecutive frames by their time difference as follows:

$$f_i(m,k) = \frac{\Delta\tilde{\varphi}(m,k)}{2\pi R} f_s \tag{3.39}$$

Where R is the hop size and f_s is the sampling frequency.

A note onset consists of an unstable transient, i.e. note at attack, followed by a stable steady state. During the steady state, the instantaneous frequency at hop m should be close to the instantaneous frequency at hop *m-1*. When this is not the case, it is an indication that an unstable event has occurred, e.g. a note onset. The difference in the instantaneous frequency between frames is proportional to the differential angle between target and current phase, $d\tilde{\varphi}$, which is calculated as follows:

$$d\tilde{\varphi} = princarg\left[\tilde{\varphi}(m,k) - 2\tilde{\varphi}(m-1,k) + \tilde{\varphi}(m-2,k)\right] \tag{3.40}$$

The transient energy, *TE*, was given by:

$$TE(n) = \sum_{k\in K} \left|X(k,nh)\right|^2 \tag{3.41}$$

where K is the set of transient frequency bins, of which a bin k is said to contain energy related to a transient if:

$$\phi(k,(n-2)h) - 2\phi(k,(n-1)h) + \phi(k,nh) < T_{tr} \tag{3.42}$$

where ϕ is the phase and T_{tr} is the transient detection threshold. It was found that some of the detection function noise in recordings with greater higher frequency content was eliminated. However, there was not a significant improvement in the uppermost band and due to the added computational cost, it was only used for the mid-frequency content.

For the lowest two sub-bands (1-2.5kHz), he proposed using a distance measure between each frame of an FFT, considering only positive values in order to reject offset detection:

$$DM = \sum_{\{k:dX_n(k)>0\}} dX_n(k)^2 \tag{3.43}$$

To detect soft as well as harder onsets, a normalization term is incorporated:

$$DM = \frac{\sum_{\{k:dX_n(k)>0\}} dX_n(k)^2}{\sum_{k=1}^{N/2} |X(k,(n-1)h)|^2} \tag{3.44}$$

where N is the FFT array length. This detection function obtains the average increase in energy per frame. Both detection methods are used on the second lowest band as it contains both energy bursts and slowly decaying pitch information.

The detection function from each band took the form of an energy / distance measure over time. Onsets were detected when the exponential weighting function given below was greater than the threshold:

$$ons(n) = SE(n) - \sum_{a=1}^{A} \frac{SE(n-a)}{a} \tag{3.45}$$

where the transient energy term, TE, or the distance measure, DM, are substituted for the energy term, SE, where required. This detection function gave most emphasis to recent frame values but still allowed previous frames have an effect, acting like a low pass filter with very smooth roll-off, where the integer a is equivalent to the filter coefficients.

The threshold was set automatically using the statistical properties of the detection function. The probability distribution function (pdf) of the detection function was estimated and the second derivative of this was then calculated. The threshold was chosen at the point where data is most likely to be an onset, which was found to occur at the maximum of the second derivative. All values in the detection function, which are equal to or greater than this, should be an onset.

The onsets were chosen from each sub-band individually and the results combined. This caused multiple onsets at the same location, which was solved by applying a 50ms window and choosing the most prominent onset. Onsets in the higher bands were always given precedence as the higher bands have better time resolution and the lower bands have better frequency resolution. The system was tested on a wide range of signals and yielded good results.

[Duxbury '03a] used the system described above in [Duxbury '02] to locate transients for improved segmentation in audio time scaling. Time scaling means to stretch or compress a piece of music without altering the pitch. Once a transient was located, the remaining steady state region of the note was stretched or compressed with excellent results being achieved for a range of signals.

[Bello '03] presented a phase-based onset detector. A transient / steady state separation method was used and the resulting data analysed using statistical methods. Phase-vocoder theory, as described on page 46 from equations (3.38) to (3.40), was used to obtain differential angles.

The pdfs of the differential angles are observed, frame-by-frame, for all k. It was discovered that when a transient occurs, the distribution becomes dispersed along the phase range. Immediately afterwards, when entering the steady state, the sharpness and height of the distribution increase.

Three methods were used to measure the spread of the distribution, the standard deviation, interquartile range (IQR) and kurtosis. IQR is a measure of spread or dispersion. It is the difference between the 75^{th} (Q3) and 25^{th} (Q1) percentiles, i.e. Q3-Q1. IQR performed better than the standard deviation as it is less sensitive to small variations in the distributions spread. It was discovered that in real recordings, phase

misalignment between partials of a note caused the differential angular distribution to spread. This is critical when notes evolve for a long time. Kurtosis is a measure of the shape of the distribution, increasing for sharp distributions and decreasing for flat distributions. The Fisher kurtosis, a common implementation, is measured as follows:

$$\gamma_2 = \frac{\mu_4(\mu)}{(\mu_2(\mu))^2} - 3 = \frac{\mu_4(\mu)}{\sigma^4} - 3 \qquad (3.46)$$

where $\mu_4(\mu)$ denotes the fourth central moment of the data and σ is the standard deviation. A peak in the kurtosis profile is an indication of when the steady state of a note occurs. The onset time of this note can be found by locating the closest preceding peak in the IQR profile.

A dynamic threshold was used to select peaks. This was calculated as the weighted median of a H-length section of the kurtosis around the corresponding frame:

$$\delta_t(m) = C_t median\gamma_2(k_m), k_m \in \left[m - \frac{H}{2}, m + \frac{H}{2}\right] \qquad (3.47)$$

where C_t is a predefined weighting value and k_m the analysis window.

The system was tested on a small database of commercial recordings using different values of C_t. They decided that an optimal value of C_t gave good detection rates of 80-90% at a cost of around a 10% rate of false positives.

[Duxbury '03b] used a combination of phase and energy based approaches to detect onsets. The histograms of phase deviation and energy difference have a similar distribution at both transient and steady state regions. Each vector of the energy difference and phase deviation were considered as data sets X_e and X_p where each x $\in$ X is a bin value and f(x) is the pdf of X. The curve describing the spread of f(x) over time became noisy when X contained few points. To overcome this, the mean of the distribution of the absolute values of X across time, $\eta(n)$, was calculated as follows:

$$\eta(n) = mean\left(f_n(|x|)\right) \qquad (3.48)$$

$\eta(n)$ yielded a sharp profile which maximises the spread of the distribution. The distributions of the phase deviation and energy difference were measured separately and the results combined by multiplying both spreads as follows:

$$\eta_T(n) = \eta_e(n) \times \eta_p(n) \tag{3.49}$$

Onset peaks were selected using equation (3.47) above, substituting $\eta_T(n)$ for γ_2.

A database of musical recordings including three solo instruments and a group recording were used to test the system. The system was robust to the signal type, which is not the case when using either detection method independently. Phase based approaches tend to be distorted by noise whereas soft onsets pose a problem to energy based approaches. The combined approach has the advantage of greatly reducing instabilities in either approach. It also yields sharper peaks at onsets, resulting in more accurate detection.

[Duxbury '03c] presented a novel method for onset detection by combining energy and phase based approaches in the complex domain. It builds upon the previous approach, described in [Duxbury '03b]. He observed that onset detection could be divided into two parts:

1. Onset Detection
2. Peak picking

The first is concerned with locating onset transients whereas the second must determine which of these onset transients are note onsets. This approach dealt mainly with the onset detection stage.

By inspecting changes in both the frequency and amplitude of audio signals, onset transients can be located. However, the effects of both variables can be simultaneously considered by predicting values in the complex domain. The target value for the k^{th} FFT bin is obtained as follows:

$$\hat{S}_k(m) = \hat{R}_k(m) e^{j\hat{\phi}_k(m)} \tag{3.50}$$

where $\hat{R}_k$ corresponds to the magnitude of the previous bin, $|S(m-1)|$ i.e. R_{k-1}, and $\hat{\phi}_k$ is the expected phase. The measured value for the same bin is given as follows:

$$S_k(m) = R_k(m)e^{j\phi_k(m)} \tag{3.51}$$

where R_k and ϕ_k are the measured magnitude and phase for the k^{th} bin of the current STFT frame. Using the Euclidean distance between the target and measured vectors, the stationarity of the kth bin can be found:

$$\Gamma_k(m) = \left\{\left[\Re(\hat{S}_k(m)) - \Re(S_k(m))\right]^2 + \left[\Im(\hat{S}_k(m)) - \Im(s_k(m))\right]^2\right\}^{\frac{1}{2}} \tag{3.52}$$

where $\Re$ and $\Im$ are the real and imaginary parts. By summing these values over all k, a frame-by-frame detection function can be obtained as follows:

$$\eta(m) = \sum_{k=1}^{K} \Gamma_k(m) \tag{3.53}$$

To simplify equation (3.50), they mapped $\hat{S}_k(m)$ onto the real axis thus forcing $\hat{\phi}_k(m)$=0. $S_k(m)$ can then be represented using the phase deviation, $d\tilde{\varphi}$, from equation (3.40):

$$S_k(m) = R_k e^{jd\tilde{\varphi}_k(m)} \tag{3.54}$$

Applying these conditions to equation (3.47) and simplifying gives:

$$\Gamma_k(m) = \left\{\hat{R}_k(m)^2 + R_k(m)^2 - 2\hat{R}_k(m)R_k(m)\cos d\tilde{\varphi}_k(m)\right\}^{\frac{1}{2}} \tag{3.55}$$

In the case where $d\tilde{\varphi}_k(m) = 0$:

$$\Gamma_k(m) = \hat{R}_k(m) - R_k(m) \tag{3.56}$$

which is equivalent to a basic amplitude difference measure for the kth bin. They found that the resulting detection function gave sharp peaks at points of poor

stationarity and was less noisy than using either an amplitude or phase based approach, thus simplifying the peak picking task.

Onset peaks were selected using equation (3.47) above. The system was tested on a wide range of polyphonic signals and the results were compared with those achieved using separate phase and energy based techniques and the same peak picking method. The complex domain method gave the best performance with an average of 95% good detections for 2% of false negatives.

In [Gainza '04b], a technique based on onset detection was presented to transcribe single ornaments in traditional Irish music. The tin whistle is a wind instrument and therefore produces a slow onset, which is more difficult to detect than a fast onset. In its most common mode, the range of this instrument is fourteen notes, from D4 to B5. The tin whistle is a transposing instrument. This means that when a musician plays D4 from a musical score, the pitch of the note played is actually D5. Many types of ornamentation are used in traditional Irish music and this paper dealt with the two single note ornaments used by the tin whistle, the cut and the strike. The cut is pitched above the parent note and the strike below. More information on ornamentation can be found in chapter 2.

Firstly, a time-frequency analysis of the signal was obtained using the STFT. The STFT representation of the signal was then split into 14 bands, one for each note in the range. The average energy envelope, E_{av}, of each band was calculated using:

$$E_{av(i,n)} = \sum_{k=1}^{l_i} \left\{ X_i(k,n) \right\}^2 \qquad \textbf{(3.57)}$$

where X_i contains the frequency bins associated with band i, k is the k^{th} frequency bin in X_i and l_i is the number of frequency bins in band i. This was convolved with a Half Hanning window. Both of these operations were carried out to smooth the signal. The first order difference function of the energy envelope for each band was then calculated to give a series of peaks. The energy increases and decreases were separated into two vectors, $D_{E(i,n)}$ and $D_{D(i,n)}$, which were inspected for onset and offset peaks that reached a predetermined band dependent threshold.

It has been noted that the blowing pressure a musician must apply to play each note on the tin whistle increases logarithmically with frequency [Martin '94]. As a result of this, the threshold, T, for each band, i, was obtained as shown in the following equation:

$$T_i = T * 2^{\frac{s}{12}} \tag{3.58}$$

Where T was a known threshold for a band, which is separated from band i by s semi-tones. A possible onset candidate was detected if:

$$D_{E(i,n)} = E_{(i,n)} - E_{(i,n-1)} \geq T_i \tag{3.59}$$

and an offset candidate when:

$$D_{D(i,n)} = E_{(i,n)} - E_{(i,n-1)} < -T_i \tag{3.60}$$

Each onset candidate was matched with the next offset candidate to form audio segments. Segments that were shorter than a time threshold were considered to be ornaments. A cut was detected if the ornament pitch was higher than the parent note and a strike was detected when the pitch was lower than the parent note. Since an ornament occurs on the beat of its parent note, the ornament onset time was also the onset time of its parent note. The system was tested on three recorded tin whistle tunes with high accuracy rates for ornamentation detection.

3.3 Chapter Summary

During this review of onset detectors, several techniques were discussed. These included energy and phase based approaches and combinations of the two. One of the most challenging aspects of an onset detector appears to be setting the threshold above which onsets are detected. Choosing a threshold that is too low results in unwanted spurious onsets, while one that is too high means that softer onsets will often be missed.

A common approach encountered during the literature review was to split the audio signal into a number of frequency bands and analyse each one individually before

combining the results at the end. This seems like a good approach and one that would be suitable for the proposed onset detection/transcription system.

4. Onset, Ornament Detection and Music Transcription

This chapter is divided into three different sections. Section 4.1 Technique Analysis contains a brief analysis of some of the onset detection and music transcription techniques defined during the literature review. It gives some advantages and disadvantages of these techniques and explains why some would be more suitable than others to use in an attempt to attain note onsets, pitches and ornamentation from monophonic traditional Irish musical signals. In section 4.2, the proposed system is presented and described in detail. An overview of the proposed system is described in section 4.1.3. The description of the proposed onset detector begins in section 4.2.4 and the music transcription is explained in section 4.2.7. Section 4.3 contains the results obtained for both onset and pitch detection and ornament detection when the system was tested using real live recordings.

4.1 Technique Analysis

4.1.1 Onset Detection

The first onset detectors considered the musical signal as a whole [Chafe '85], while others [Scheirer '98], [Klapuri '99] expanded on this approach by dividing the signal into a number of frequency bands. Both energy [Masri '96] and phase based [Bello '03] approaches have been implemented; sometimes a combination of both has been used [Duxbury '03b]. There have been techniques where different detection functions have been applied to the low and high frequency bands [Duxbury '02]. Due to the success and widespread use of sub-band systems in onset detection, they appear to be a suitable approach to use in the proposed system.

4.1.2 Zero-Crossings

The zero-crossings method [Cooper '94] is a time-domain technique and is only suitable for use with monophonic signals. It involves dividing a musical signal into segments according to zero crossing points, locating similar segments and determining the distance between them. The frequency is calculated by multiplying the inverse of this distance by the sampling frequency. The robustness of this method is questionable due to the unpredictability and instability of real musical signals.

4.1.3 Autocorrelation

This is another example of a time-domain technique. Correlation is used to determine if two random processes are independent or if there is some connection between them, autocorrelation is the correlation of the waveform with itself. Autocorrelation has been successfully used as a method of pitch detection in previous monophonic transcription systems [Brown '91], [Bello '00]. As in the zero-crossing technique, it is not suitable for use in a polyphonic system as autocorrelation fuses sounds together, which prevents the separate treatment of harmonic partials. In practise, frequency based approaches have been found to be faster and more efficient [Plumbley '02].

4.1.4 Constant Q Transform

The Constant Q Transform [Brown '92] is closely related to the Fourier Transform in that it can be viewed as a bank of filters. However, it has geometrically spaced centre frequencies, which are given by:

$$f_k = f_0 . 2^{\frac{k}{b}} \tag{4.1}$$

where b is the number of filters per octave, k the octave number and f_0 the centre frequency of the lowest band. This makes the constant Q transform very useful in analysing music since by choosing appropriate values for b and f_0 the centre frequencies can be made to directly correspond to musical notes. By choosing b= 12, since there are twelve semi-tones in an octave, and f_0 to be the frequency of midi note zero means that the k^{th} constant Q bin will correspond to midi note number k. Since the constant Q uses the FFT in implementation, it is just as efficient to create an approximate constant Q by combining the filter outputs of an STFT.

4.1.5 Wavelets

While the Fourier Transform represents a signal as a set of sine waves of different frequencies, the Wavelet Transform represents a signal as a set of shifted and scaled versions of an original wavelet. A wavelet is a waveform of limited duration with an average value of zero and tends to be irregular and asymmetric. Computationally it has been found to be similar to the 'slow' Fourier transform [Sillem '00] and although

it gave better time resolution for higher frequencies than the STFT, there was not much improvement as regards frequency resolution. Overall the STFT proved to be more efficient.

4.1.6 Short Time Fourier Transform

This is an example of a frequency-based technique. The STFT is a popular choice for music transcription since it gives a frequency representation of a time domain signal. It has been used extensively throughout the Literature Review, both in music transcription [Piszczalski '77], [Martin '96a], [Klapuri '00], [Bello '00] and onset detection [Duxbury '02], [Gainza '04b], with good results being achieved. One disadvantage is the time-frequency trade off, very good frequency resolution means poor time resolution and very good time resolution indicates that frequency resolution is compromised. This can be overcome by using a short window for time resolution and a long window for frequency resolution. The STFT was the preferred analysis method for detecting note onsets, ornamentation and note pitches in the proposed system.

4.2 Proposed Energy-Based Approach

4.2.1 Note Range

The proposed system uses a sub-band approach to detect the onset of each note and ornament. Twenty-nine bands are used; one band for each of the semi-tones in the note range G3 to B5, shown below in figure 4.1. This covers the range of each of the instruments used to test the system; the fiddle, flute and tin whistle. When played in a classical style, the violin has a greater note range but after analysis of fiddle tunes it was discovered that G3 to B5 covers the notes used by a traditional musician. It is very unlikely that a note outside of this range would be played. The range of both the flute and tin whistle in traditional Irish music is D4 to B5. However, the tin whistle is a transposing instrument and so when a musician plays D4 from a musical score, the pitch of the note that is actually played is D5.

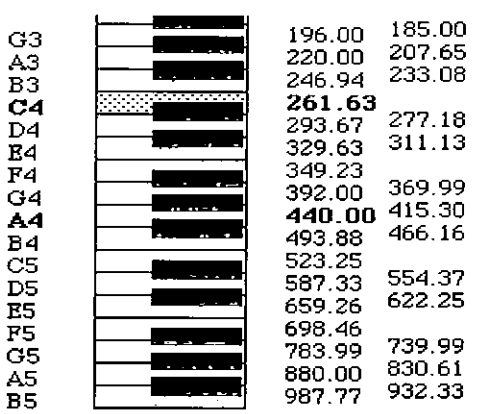

Figure 4.1: The note range of the proposed system, shown on a piano keyboard [Wolfe]

4.2.2 Initial Attempts

Initial attempts were based on those of [Piszczalski '77] and used synthesised traditional Irish tunes. The time frequency representation of the signal was obtained using the STFT:

$$X(k,n) = \sum_{m=0}^{L-1} x(m+nH)w(m)e^{-j(2\pi/N)km} \qquad (4.2)$$

where *x(m)* is the signal, *k* the frequency bin number, *n* the frame number, *H* the hop length and *w(m)* the window of length *L*.. A 1024-sample Hamming window with a 50% overlap was applied; this is based on a cosine function, gives good side-lobe attenuation and has a wide main-lobe. After experimentation it was found that a 2048-point FFT gave sufficiently good frequency resolution.

A spectrogram of one of the signals can be seen in figure 4.2, the tune is comprised of eight notes and one single note ornament. The eight notes can clearly be seen along the time axis. The ornament is located between the sixth and seventh notes, indicated by a red arrow and surrounded by a red box. Some harmonics of the first note can be seen along the frequency axis. The highest peak in each frame was located and the corresponding frequency bin number, found on the frequency axis, was retained.

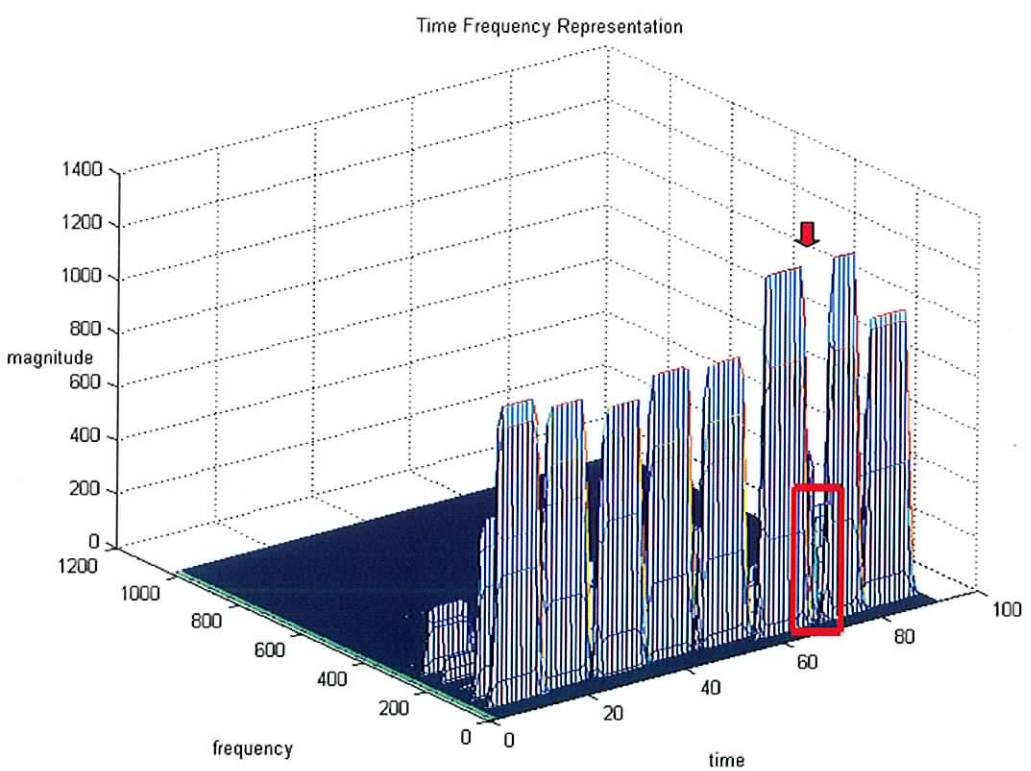

Figure 4.2: Spectrogram of tune

To locate onsets and offsets, each frequency bin number corresponding to a peak was analysed over time. An example of the amplitude envelope of the bin number of the fourth note can be seen in figure 4.3. A threshold, *T*, was set as follows:

$$T = \frac{m}{20} \tag{4.3}$$

where m was the height of the largest peak in the time evolution of a bin number. Once the amplitude rose above this threshold an onset was detected and when it dropped below the threshold an offset was detected. This value was chosen through experimentation, it was discovered that if the threshold value was too low, fluctuations in the signal were picked up as onsets. If it was too high, the duration of the note was interpreted as being shorter than it actually was. Note durations were obtained by subtracting onset times from offset times.

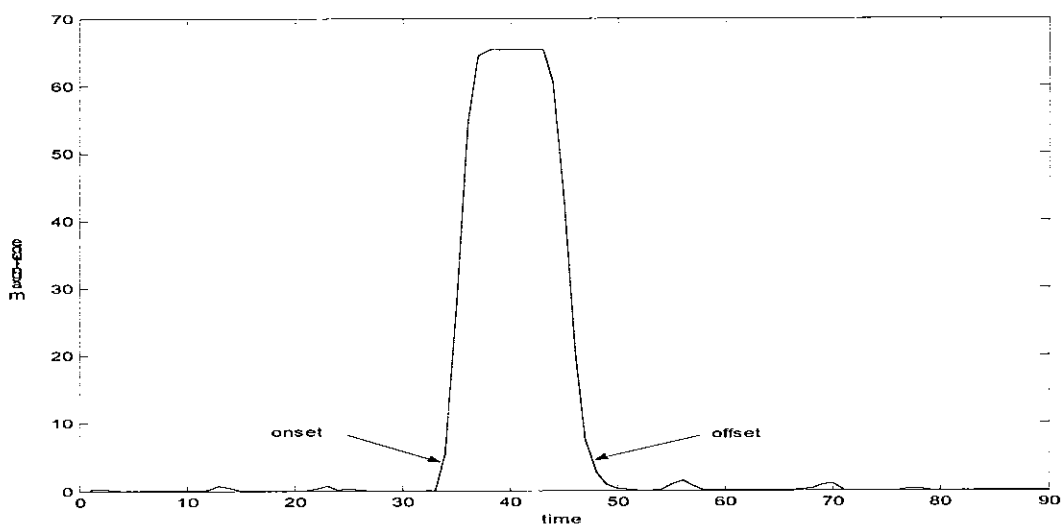

Figure 4.3: Amplitude Envelope of Frequency bin number of note

Note pitches were calculated from the frequency bin numbers as follows:

$$f = k * \left(\frac{F_s}{fft\_size} \right) \tag{4.4}$$

where f is the note frequency in Hertz, k the frequency bin number and F_s the sampling frequency.

Even with synthesised tunes this method encountered difficulties. When two notes of the same frequency were played close together, as in figure 4.4, the offset of the first note never dropped below the threshold and so the two notes were picked up as one. The duration of the ornament in the above example was detected as being the same

length as the other notes in the tune when it is in fact five times shorter. These errors are a strong indication that this method of onset detection is unsuitable. However, the results for pitch detection were encouraging with correct detections in the majority of cases.

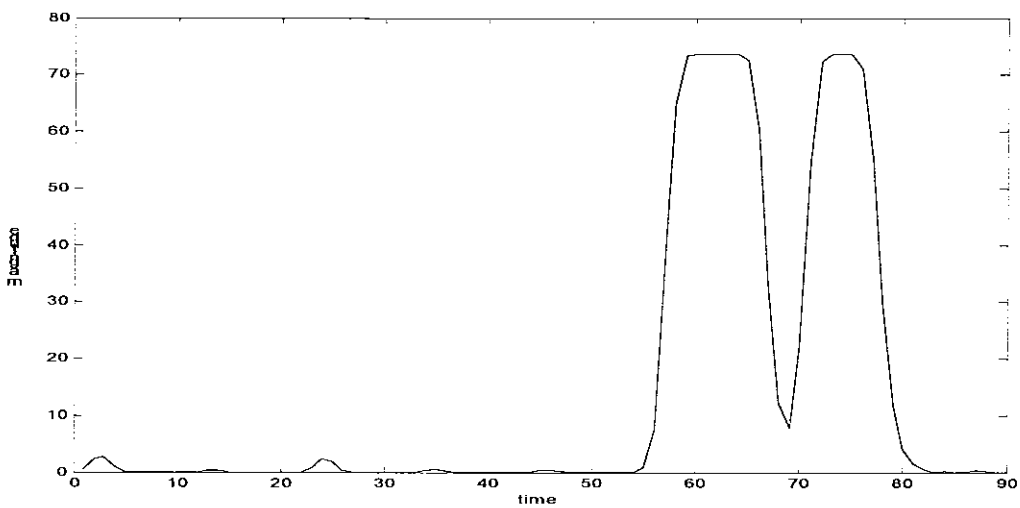

Figure 4.4: Two notes played simultaneously

4.1.3 System Overview

An overview of the proposed energy-based approach can be described as follows:

1. Obtain a time-frequency representation of the signal.
2. Split this representation into logarithmically spaced frequency bands, one for each note in the range.
3. Calculate the 1st order difference function of the energy envelope of each band.
4. Use a band dependant threshold to determine if a peak is an onset.
5. Combine bands and determine note pitches.
6. Apply music theory to determine if an onset is a note or ornament.

4.2.4 Frequency Analysis

The audio signals used had a sample rate of 44100Hz, this is the standard sampling rate for CD audio signals. Firstly a time-frequency representation of the signal is obtained, this is achieved using the STFT, equation (4.2). Each frame was created by multiplying the signal by a 1024 sample Hanning window with a hop length of 512 samples, giving a 50% overlap between frames. A 4096 point-FFT was applied to each frame.

The output of the STFT can be interpreted as a collection of uniform filter outputs. To create the bands for each note, the frames were combined using upper and lower band limits in accordance with the frequency of each note. Note semi-tones are logarithmically spaced and can be calculated using the following equation:

$$f_i = f * 2^{\frac{s}{12}} \tag{4.5}$$

where f is a reference frequency, which is separated from f_i, the frequency of semi-tone i, by s semi-tones. The frequency of each of the 29 semi-tones, from G3 at 196Hz to B5 at 987.77Hz, was converted to its frequency bin equivalent as follows:

$$k = f * \left(\frac{fft\_size}{F_s} \right) \tag{4.6}$$

where k is the frequency bin number, f the note frequency in Hertz and F_s the sampling frequency. The lower band limit of a note was set by obtaining the midpoint between the notes bin number and that of the note before it and the upper band limit by obtaining the midpoint between the notes bin number and that of the note above it. In effect, this can be viewed as an approximation of a constant Q transform. The STFT outputs have been grouped together to create a number of frequency bands whose center frequencies are logarithmically spaced.

The average energy of each band was calculated as follows:

$$E_{av(i,n)} = \sum_{k=1}^{I_i} \left\{ |X_i(k,n)|^2 \right\} \tag{4.7}$$

where X_i contains the frequency bins associated with band i, k is the k^{th} frequency bin in X_i and l_i is the number of frequency bins in band i. This gives the amplitude envelope, which contains the energy profile of the band associated with a given note. Additional smoothing is also carried out by convolving the energy envelope of each band with a half-hanning window. This performs in much the same way as the human auditory system, masking fast amplitude modulations and emphasising most recent inputs [Klapuri '99]. The first order difference function of each band is then calculated. All negative values of this function are set to zero so that only the energy increases are considered.

This results in a series of peaks in each band, which are possible note or ornament onsets. Figure 4.5 below shows the audio signal of an excerpt of a fiddle tune containing thirteen notes composed of four different pitches, along with the 1st order difference function of the note bands. Figure 4.6 shows a tin whistle tune with the 1st order difference functions of its twelve notes containing seven different pitches. Peaks are clearly visible at each note onset in both tunes. Now we are faced with the problem of determining which of these peaks are in fact note and ornament onsets, and which are merely spurious onsets which could be due to noise or amplitude modulation. This was achieved by automatically choosing a suitable threshold for each band.

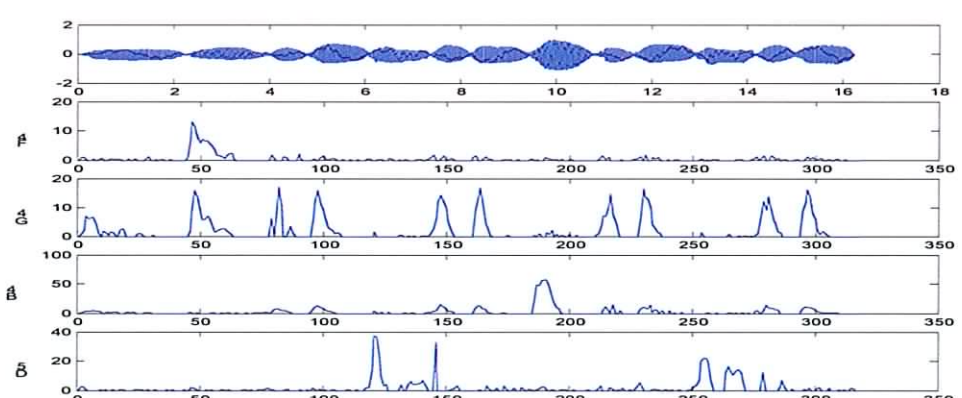

Figure 4.5: Wave file of fiddle tune, shown with the 1st order difference function of note bands

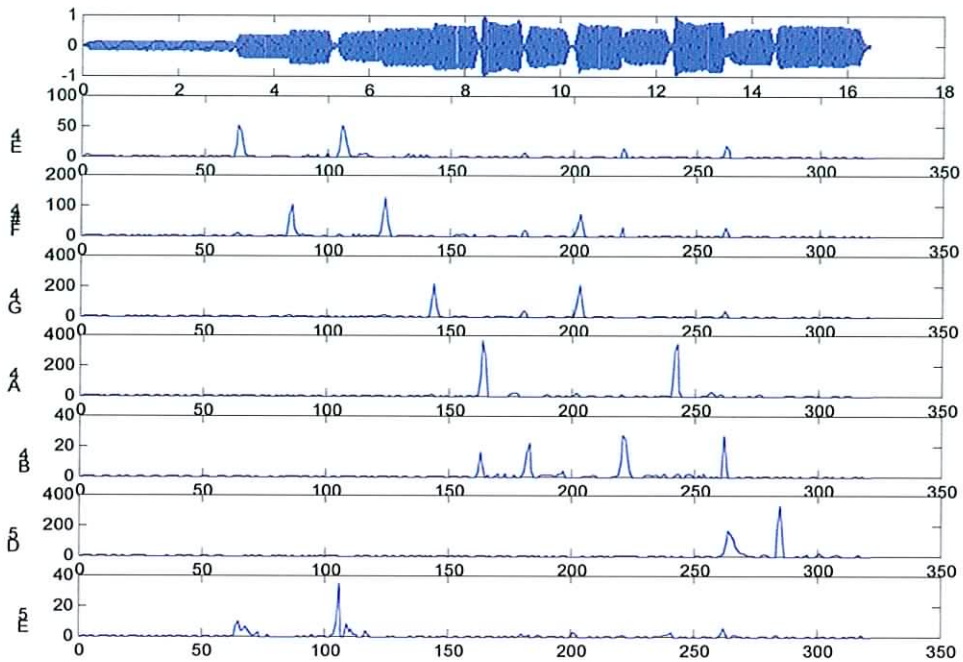

Figure 4.6: Wave file of tin whistle tune, shown with the 1st order difference function of note bands

4.2.5 Automatic Threshold

There is a certain threshold in each band where a peak equal to or greater than this value is more than likely an onset. Choosing a suitable threshold is arguably the most difficult part of onset detection. Setting the threshold too low leads to fluctuations in the signal being interpreted as notes, setting it too high and weaker onsets are missed. It is not sufficient to have one threshold to cover all bands; this is evident in figure 4.5 and particularly figure 4.6 above, where amplitude values vary greatly in each band. At the beginning, thresholds were chosen by finding the maximum amplitude value in the first order difference function of a band, and dividing this value by an integer like in equation (4.3). This proved to be an inadequate way of setting a threshold, as it was impossible to choose an appropriate divisor.

In the case of the tin whistle, it has been noted that the blowing pressure a musician must apply to play each note increases logarithmically with frequency [Martin '94]. For example, a musician would have to blow twice as hard to play a D5 as they would

to play a D4. As a result of this, Gainza, Kelleher, et al. determined that a suitable threshold, T, for each band, i, when transcribing the tin whistle can be obtained as shown in the following equation:

$$T_i = T * 2^{\frac{s}{12}} \quad \textbf{(4.8)}$$

Where T is a known threshold for a band, which is separated from band i by s semi-tones [Gainza '04a] and each band contains the frequency range for one semi-tone.

The bands were then recombined and compared in order to determine possible note onsets. When two or more peaks are contained in the same, previous or next frame, i.e. within 46ms, the strongest is retained as the possible onset candidate. This is because it is not possible for the human auditory system to distinguish between two onsets occurring within a 50ms timeframe and the most intense onset will more than likely mask any others. Next, a 46ms window was applied to each onset candidate. In the case where there is one onset in the window, the frame number is retained as the note onset time and the band number is retained as the note pitch. When two peaks occur in the window an articulation has occurred, the frame number of the first peak is retained as the note onset time and the band number of the second peak as the note pitch.

Two tin whistle tunes from [Larsen '03] were used to test the system, a 17 second excerpt from "The boys of Ballisodare" (p. 134) and a 16 second excerpt from "Bantry Bay" (p. 152). The results were compared with those using Klapuri's system from [Klapuri '99] and are shown in Table 1. Comments on each tune follow the table. The percentage of correct onset detections were calculated using the following equation:

$$\%correct = \frac{no\_of\_notes - (undetected + spurious)}{no\_of\_notes} \times 100 \quad \textbf{(4.9)}$$

Tune	System	Undetected	Spurious	% Correct
1	Tin Whistle	0/50	3	94
1	Klapuri	4/50	3	93.4
2	Tin Whistle	0/42	0	100
2	Klapuri	2/42	2	85.7

Table 1: Tin whistle onset detection comparison

1. The three spurious onsets were detected when a stepwise descending note was played with a cut. This is the most complex type of cut to play and as it was played for longer than the set ornamentation duration threshold and was considered to be an independent onset by the system.
2. All note onsets were detected correctly.

The results for pitch detection can be seen in Table 2. Comments on each tune follow the table.

Tune	Correct pitch detections (%)
1	94
2	97.6

Table 2: Pitch transcription results

1. The pitch for each detected onset was transcribed correctly.
2. One onset was transcribed incorrectly when a strike was picked up as its parent note.

As can be seen from the results, the tin whistle onset detection system performed better than that of Klapuri. The band dependent threshold was adequate at dealing with strong signal modulations. Klapuri's system had some problems detecting fast changes between notes in the same band. Customising the system according to the characteristics of the instrument improved the onset detection accuracy. However, this approach is unsuitable for use with the fiddle as it is a string and not a wind instrument. As a result of this, an alternative method had to be used to determine thresholds, which was also suitable for use with the fiddle.

A threshold can be automatically chosen for each band by analysing the statistics of their energy envelopes [Duxbury '02]. Considering a histogram of the amplitude envelope of a band, the ideal threshold is at the point where the data is more likely to be an onset. This threshold was determined as follows; firstly, the probability distribution function (pdf) of the energy envelope was estimated using a smoothed histogram, followed by obtaining the second derivative of the pdf. The point where all greater values in the first order difference function should represent note onsets is where the curve in this second derivative takes the characteristic of the transient component.

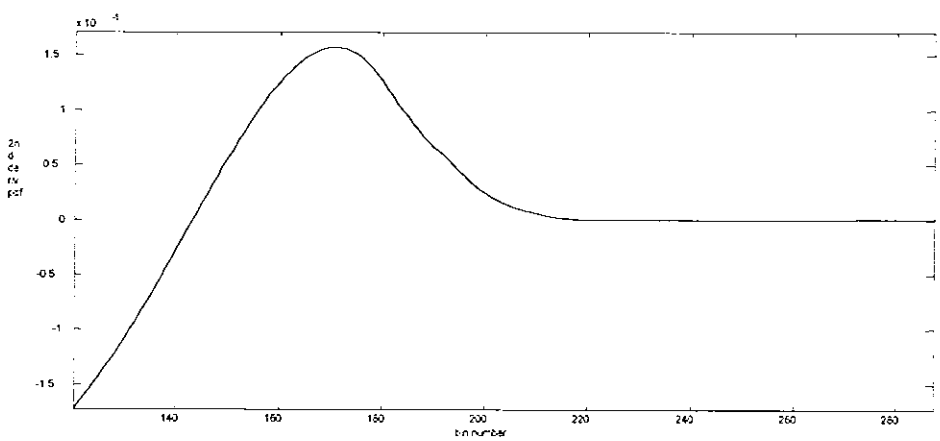

Figure 4.7: Second derivative of frequency band energy envelope

This point can be seen as the peak in figure 4.7. Duxbury found that a peak in the first order difference function whose amplitude is greater than or equal to this maximum is probably an onset. Peaks with amplitudes less than this threshold are likely to be due to noise and fluctuations in the amplitude envelope of the signal. Therefore this maximum is chosen as the band threshold. The peak in each frame is compared with the peaks, if any, in the previous and next frames. The peak of greatest amplitude is kept and the other frames are set to zero. Peaks that are greater than the band threshold are retained as possible onset locations. Analysis of this onset detector is carried out in section 4.3.

4.2.6 Band Combining

Every band is compared with each of the other 28 bands to determine possible onset locations. A peak in the current frame of the current band is compared with the previous, current and next frame of all other bands. When there are no peaks in the other bands, the current frame is chosen as the possible onset location. In cases where there are peaks in the other bands, the peak with the greatest magnitude is chosen to be the possible onset location and the same frame in the other 28 bands is set to zero. This results in no more than one peak per frame across the 29 bands. All bands are then combined to give a series of peak locations, which are possible onset candidates, in time order.

To give a more accurate onset time, the onset locations are adjusted to be three frames before those detected. This is necessary as there is a gradual slope to each peak maximum. We have observed in practice that this operation gives a similar result to obtaining the relative difference function of the signal as used by Klapuri to get onset times in [Klapuri '99] but is simpler to implement. This slope is clearly seen in figure 4.8, which is the first order difference function of the energy envelope of a fiddle note. The onset occurs at frame 1 but the maximum amplitude is not reached until frame 4.

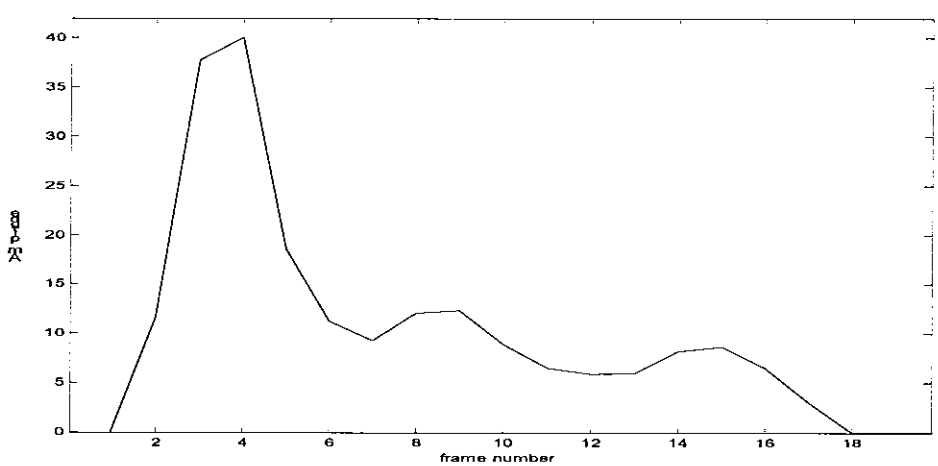

Figure 4.8: Fiddle note onset peak showing gradual slope to maximum

Onsets are located in the original wave file as follows:

$$onset\_location = frame\_no.* H \tag{4.10}$$

where H is the hop length used in the original STFT, 512 samples in this case.

A second approximation of the onset locations is obtained in a simple summation of the energies in each frame of an STFT of the original signal:

$$E(n) = \sum_{k=1}^{L-1} \left\{ \left| F(k,n) \right| \right\} \tag{4.11}$$

where $F(k)$ contains the frequency bins, k, associated with frame n and L is the window length used in the STFT. This energy summation is shown in figure 4.9. The onset times of these peaks are found by locating the minima in this summation, as there is a gradual slope from zero to the maximum value of each peak, similar to that described above in the sub-band case. Therefore taking the location of the minima as opposed to the maxima gives a more accurate result.

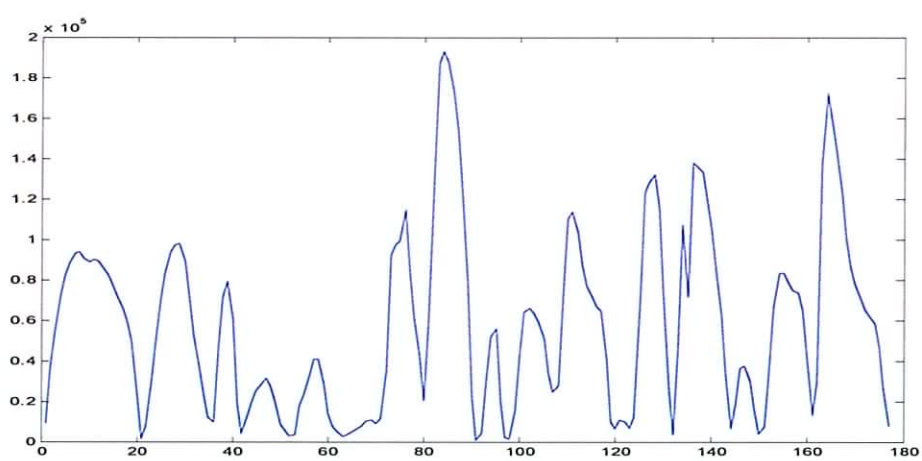

Figure 4.9: Summation of the energies in each STFT frame of a fiddle tune

Although the sub-band procedure described above does an adequate job detecting the note onsets of a tune, this energy summation is intended to pick up any onsets that may have been missed and combining both techniques results in a more robust system. It should be noted that the energy summation does not detect ornaments and is merely a way of eliminating gross errors in the detection system. It is of some help

in the case of the fiddle and tin whistle but is of little or no benefit for the flute. This is because the energy summation waveform for the flute has unpronounced minima, as shown below in figure 4.10. To detect the peak maxima would require setting a threshold and this would be extremely difficult due to the constant signal energy modulations.

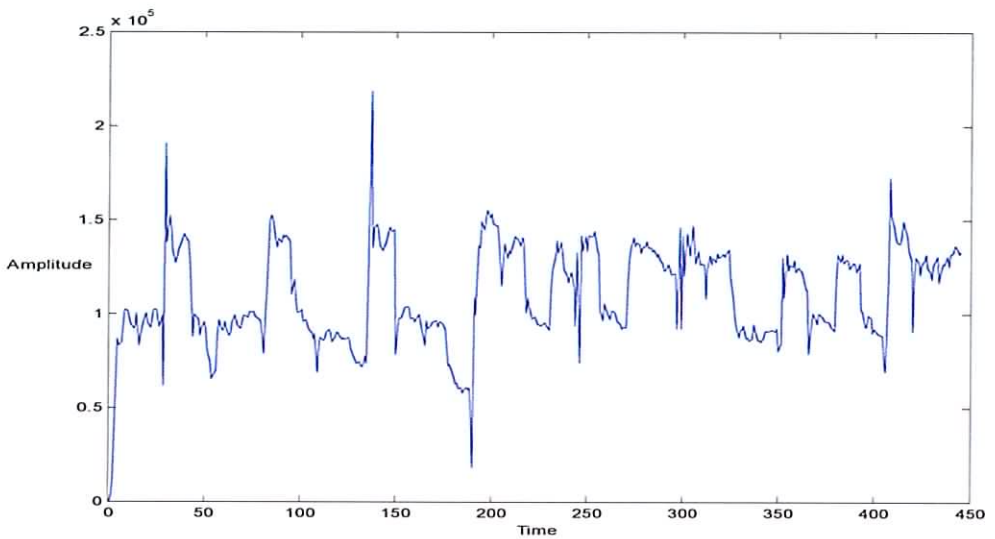

Figure 4.10: Summation of the energies in each STFT frame of a flute tune

Figure 4.11 shows a situation where the energy summation proves useful in the case of a fiddle tune; the onsets are denoted by the red lines. It has picked up the 4th onset, which was missed by the sub-band approach. Both sets of onsets are then compared with multiples at the same location eliminated.

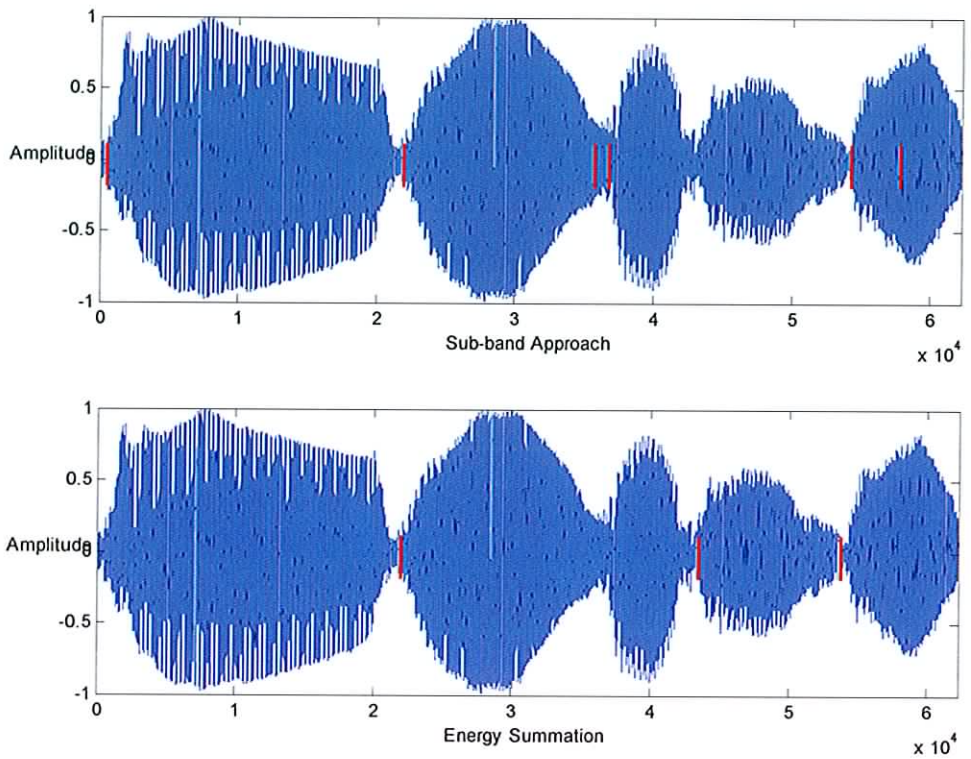

Figure 4.11: Comparison of both onset detection technique results prior to combining

A novel onset detector has been presented, consisting of the best elements of other systems combined. The sub-band approach from [Klapuri '99] was used, although each frequency band in his approach contained at least four notes whereas the onset detector proposed above used one frequency band for each note. The amplitude envelope and first order difference function of each frequency band was calculated, just as in Klapuris system. A band dependent threshold was automatically chosen for each band by analysing the statistics of its energy envelope as in [Duxbury '02]. There was a slight difference in that Duxbury analysed the statistics of his detection function whereas the above system analysed the amplitude envelope statistics. A novel idea in the proposed system was the addition of an efficient double check for missed onsets in the STFT energy summation, the advantages of which can be seen in figure 4.11 above.

4.2.7 Note Pitch and Ornament Identification

While the sub-band approach is a reliable method of determining the onset locations of a note, it is not the case when determining note pitches. In some cases the onset peak occurred in a band a semi-tone or two above or below the actual pitch of a note. This is due to the time-frequency trade off of using the STFT. For good time resolution, required for detecting onsets, a short analysis window is used. Good frequency resolution is required for pitch detection so a longer analysis window must be used. For this reason, the band frequency is not relied upon as being equal to the fundamental frequency of a note.

The fundamental frequencies of the notes are determined by obtaining excerpts from the waveform and carrying out frequency analysis on each excerpt. The possible onsets are located in the wave file using Equation (4.10), and analysis is carried out once a third of each note duration has passed. In the case of very short notes, i.e. possible ornaments, analysis occurs three samples into the note. The reason pitch estimation does not take place immediately when the note begins is because a more accurate result can be obtained by analysing the note after it has entered its steady state where the pitch of the note is more clearly established.

Once the analysis starting point has been ascertained a 4096-sample window is applied to the signal, beginning at this starting point. This is four times longer than the analysis window used during onset detection and should give much better frequency resolution. A 4096-point FFT is then carried out on this section of the wave file The fundamental frequencies of the considered note range, from G3 at 196Hz to B5 at 987.77Hz, are found in the frequency bins numbered 1 to 103. Therefore the location of the peak within this frequency bin range is found for each onset. These frequency bin numbers are converted into their corresponding frequency in Hertz using Equation (4.4) above. We now have a series of possible onset locations and their corresponding pitches and all that must be determined is whether these onsets are note or ornament onsets, or if they are spurious.

Examples of four types of ornament, along with their respective parent notes are shown in figure 4.12 below, onsets are indicated from above using black arrows. Example (a) is a cut played by a flute, (b) is a strike played by a tin whistle, (c) is a

double cut played by a fiddle and (d) is a long roll played by a tin whistle. Examples (a) and (b) are single note ornaments, (c) is a double note ornament and (d) is a multi-note ornament consisting of parent note-cut-parent note-strike-parent note. This figure gives a good indication of how short an ornament really is, the ornaments are roughly seven times shorter than the parent note, which in itself is only 320ms. In the case of the double cut, this means that each of its individual 'notes' are fourteen times shorter than the parent note.

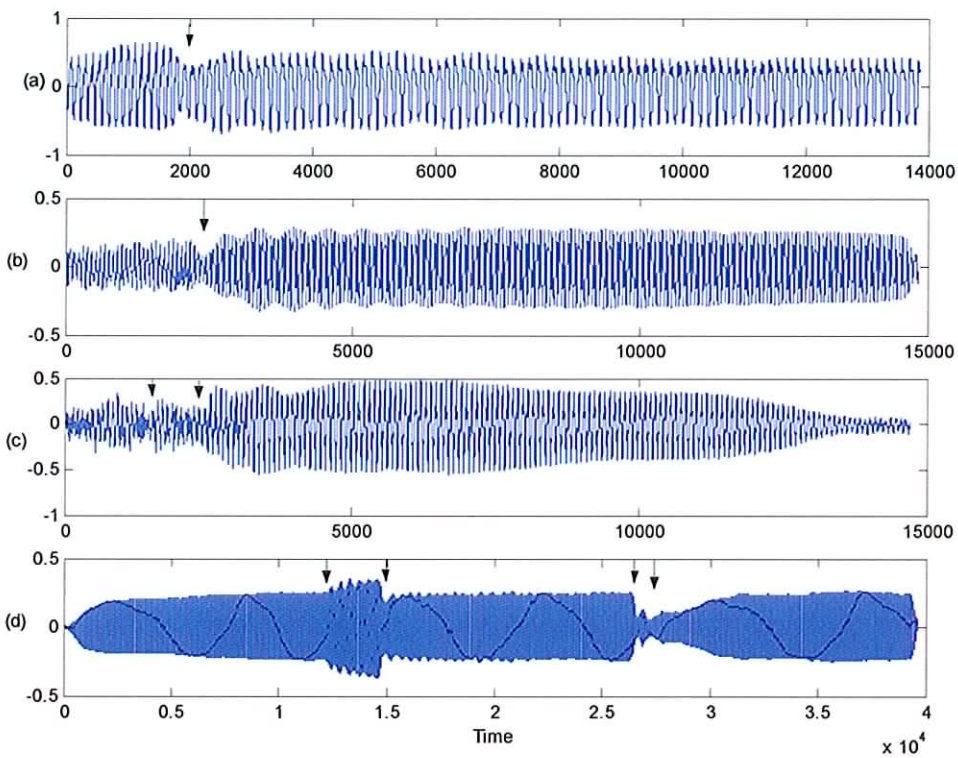

Figure 4.12: Examples of ornament plots; (a) is a cut played by a flute, (b) a strike played by a tin whistle, (c) a double cut played by a fiddle and (d) a roll played by a tin whistle

The system dealt with the identification of three different ornaments, the cut, strike and double cut. Multi-note ornaments such as the roll were beyond the scope of this project.

To establish whether an onset is a note, ornament or spurious, the distance between each consecutive onset is obtained. Considering the scenario of two onsets where

onset x occurs before onset y, a number of outcomes are possible. The two note onsets are expected to be a certain minimum distance, D, apart, i.e.:

$$y - x \geq D \tag{4.12}$$

At first, this distance D was calculated by obtaining a fractional value of the average distance between each note in a tune. After experimentation, the average note length was divided by three. Results using this method were not consistent and there were many times when regular notes were detected as ornaments and ornaments detected as regular notes so another method of setting the ornamentation duration was sought.

After analysis of the database of traditional Irish tunes used in the study, it was discovered that no regular note was shorter than 136ms, which is equivalent to 11 frames. Consequently if D is greater than 11 frames x is a regular note onset, onset y is compared to the next onset and so on. In the case where D is less than 11 frames, if x and y have the same pitch, x is again a regular note onset and this time onset y is spurious. When x and y are of different pitches, x is an ornament onset, onset y is compared to the next onset and so on. If $x > y$ then x is a cut and when $x < y$, x is a strike. The note frequencies were then assigned their equivalent note name, i.e. G3 to B5. If an onset was that of an ornament, a cut was represented by a * and a strike by a ^. The result of this process is a series of note and ornament onsets of varying frequencies, which should correspond with those contained in the original wave file.

An example of correct onset and pitch detection is shown below in figure 4.13. It is the same tune segment used in figure 4.11 above. The note names are displayed at the onset locations.

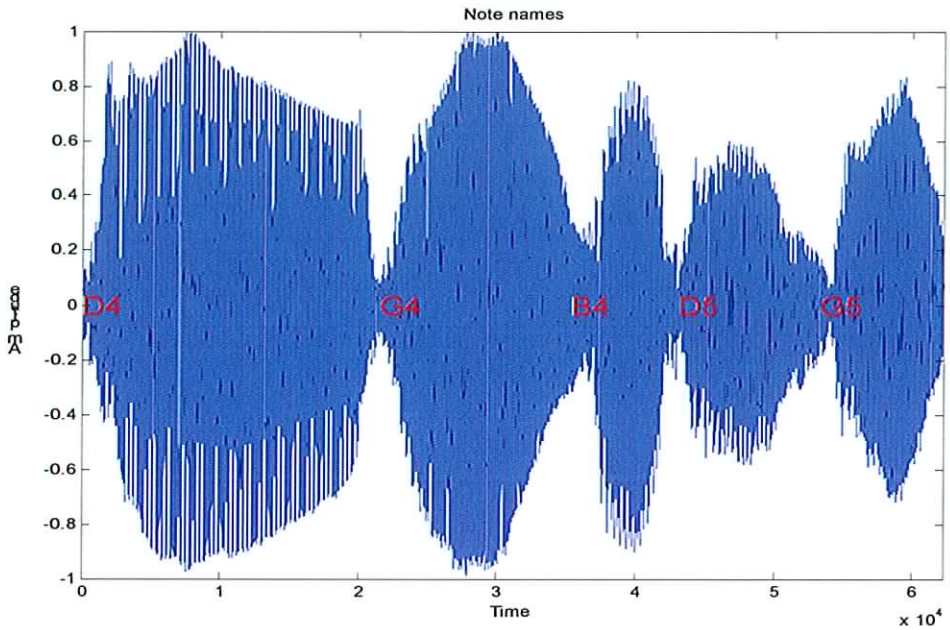

Figure 4.13: Example of correct onset and pitch detection

4.3 Results

In [Kelleher '05], the system described in Section 4.2 above was used to detect note and ornament onsets in fiddle tunes. When detecting ornaments if x has a greater pitch than y, and y is a regular note then x is a cut and y the onset of its parent note. Another possible scenario is when the pitch of x is less than y and y is also an ornament whose pitch is greater than the note, z, which follows it. If z is a regular note, x and y are the onsets of a double cut and z the onset of its parent note.

Real live recordings from Matt Cranitch's *Irish Fiddle Book* [Cranitch '01] were used to test the system with ten excerpts from ten different tunes selected. Appendix A contains an index of where to find the tunes, listed in Table 3. All but three of these excerpts contained single or double note ornaments and were, on average, ten seconds long. As these were real recordings played on a fiddle, it was found upon very close inspection of the waveform that what was played did not always exactly match the notation, particularly in the case of ornamentation. This was picked up by the detection system and was not noticeable to the untrained listener. On that basis it was thought it would be more accurate to compare the detection results to what was actually played, as transcribed by an expert listener, rather than the given notation. The results for onset detection of regular notes can be seen below in Table 3. The system detected regular onsets quite well with an average accuracy of 91% in the tunes that were tested. Percentages for accurate regular onset detection were calculated as follows:

$$\%correct = \frac{no\_of\_notes - (un + spur + orn)}{no\_of\_notes} \times 100 \qquad (4.13)$$

where *no_of_notes* is the number of notes in the tune, *un* undetected notes, *spur* spurious onsets and *orn* notes that were detected as ornaments. Pitch detection results were also good with an average accuracy of 89%. Percentages for accurate pitch detection were calculated using the following equation:

$$\%correct = \frac{no\_of\_notes - (un + spur + orn + pitch)}{no\_of\_notes} \times 100 \qquad (4.14)$$

where *pitch* is the number of pitches detected incorrectly in the tune. Comments on each tune follow the table.

Tune	Undetected	Wrong Pitch	Spurious	Detected as ornament	% Correct Pitch Detection	% Correct Onset Detection
1. Give us a drink of water	0	0	0	0	100	100
2. The Connachtman's Rambles	1/35	0	0	0	97	97
3. Dalaigh's Polka	1/34	0	0	1/34	94	94
4. The Humours of Carrigaholt	4/26	0	0	0	85	85
5. Denis Murphy's Slide	5/38	0	0	0	87	87
6. Cronin's Hornpipe	1/43	2/43	0	0	93	97
7. The Peeler's Jacket	2/34	0	0	0	94	94
8. The Hag's Purse	0	1/38	0	0	97	100
9. The top of Maol	0	0	0	4/33	88	88
10. The Lakes of Sligo	4/40	0	0	4/80	80	80
11. The Scartaglen Slide	1/42	6/42	0	6/42	69	83

Table 3: Onset detection results for fiddle

1. All note onsets and pitches were detected correctly.
2. The undetected note was the first of two A4s played in succession and had a very weak onset.
3. There was an unclear transition to the undetected note from the note preceding it. One note was detected as a double cut, this may be because this note was a semi quaver, half the length of the other notes in the tune, and there was a short change in pitch to A7 at the beginning of this note.
4. The four undetected notes had very weak onsets.
5. The five undetected notes had weak onsets, particularly those of the 1st and last missed notes.
6. Two notes were detected as being an octave too high. In both cases G5 was detected instead of G4, and a spurious strike was detected immediately before each note. This occurred because of a brief pitch change from G4 to G5 during the notes. In the second case the pitch change from G5 back to G4 at the end of the tune was detected as a strike before the note following it. The undetected note had a weak onset.
7. Both of the undetected notes had weak onsets.
8. The pitch of one note was detected a semi-tone high as G4 instead of F#4.
9. The four note onsets interpreted as ornaments were semi-quavers.
10. Two of the undetected notes were semi-quavers and one of the remaining undetected notes had a weak onset. The four notes interpreted as ornaments were semi-quavers.
11. The undetected note had a very weak onset. Of the six pitch errors, four were a semi-tone higher than they should have been and were always E4 interpreted as F4. As there were no other E4s played in the tune, this may have been due to the tuning of the fiddle. The other two were octave errors, G5 instead of G4 and in the first of these a strike with a pitch of G4 was detected immediately before. The six notes that were incorrectly interpreted as ornaments occurred during fast transitions between notes.

Results for ornament detection are in Table 4, where c, s and dc represent a cut, strike and double cut respectively. The double cut posed most problems to the system. This is a double note ornament and the system often failed to detect the onset of the second note, as it was generally less than 10ms long. The first note of a double cut is usually

of the same frequency as its parent note, which meant that it would be treated as a spurious onset in the selection process if the onset of the second note of the double cut was not detected. In other cases, a double cut was detected as a cut. The average accuracy was low for cuts at 38% and double cuts at 25%. The one strike was detected giving an accuracy of 100% Percentages were calculated as follows:

$$\%correct = \frac{detcted - spurious}{no\_of\_ornaments} \times 100 \quad \textbf{(4.15)}$$

where *detected* is the number of ornaments detected correctly, *spurious* is the number of spurious ornaments that were detected and *no_of_ornaments* is the number of ornaments in the tune.

Tune	Ornament Types	Detected	Spurious	% Correct Double Cut	% Correct Strike	% Correct Cut	% Correct Ornament Detection
1. Give us a drink of water	1 dc	1 dc	0	100	-	-	100
2. The Connachtman's Rambles	2 c 1 s	2 c 1 s	0	-	100	100	100
3. Dalaigh's Polka	6 dc	3c	1 dc	0	-	-	0
4. The Humours of Carrigaholt	1 c	0	1s	-	-	0	0
5. Denis Murphy's Slide	1dc	0	0	0	-	-	0
6. Cronin's Hornpipe	4c	2c	3 s	-	-	50	50
7. The Peeler's Jacket	1 c	0	1s	-	-	0	0
8. The Hag's Purse	2 dc	0	0	0	-	-	0

Table 4: Ornamentation results for fiddle

1. The double cut was detected correctly.
2. Both ornaments were detected correctly.
3. The 'notes' of the double cuts were so short that the 1st was not detected in all cases and the 2nd in the 1st, 4th and 5th. The reason for the spurious double cut is explained above in the comments on Table 3 above.
4. The cut was undetected as it was 7ms long, less than 1/3 of a frame. A spurious strike was detected where a short B4 had been played.
5. The second 'note' of the double cut was 10ms long and undetected.
6. The spurious strikes that were detected have been explained in the comments on Table 3 above. Neither of the first two cuts were detected.

7. The cut was undetected and a spurious strike occurred where there was a change in pitch towards the end of a note.
8. Neither double cut was detected. The 1st was a very 'blurred' cut and the onset of the second 'note' of the 2nd cut was not detected and so its 1st 'note' was treated as a spurious onset.

A representation of the double cut played in 'Give us a drink of water,' tune 1 in Table 4 above, is shown in figure 4.14 below. The double cut occurs immediately before C5, it's strike and cut are is illustrated using the symbols ^ and * respectively.

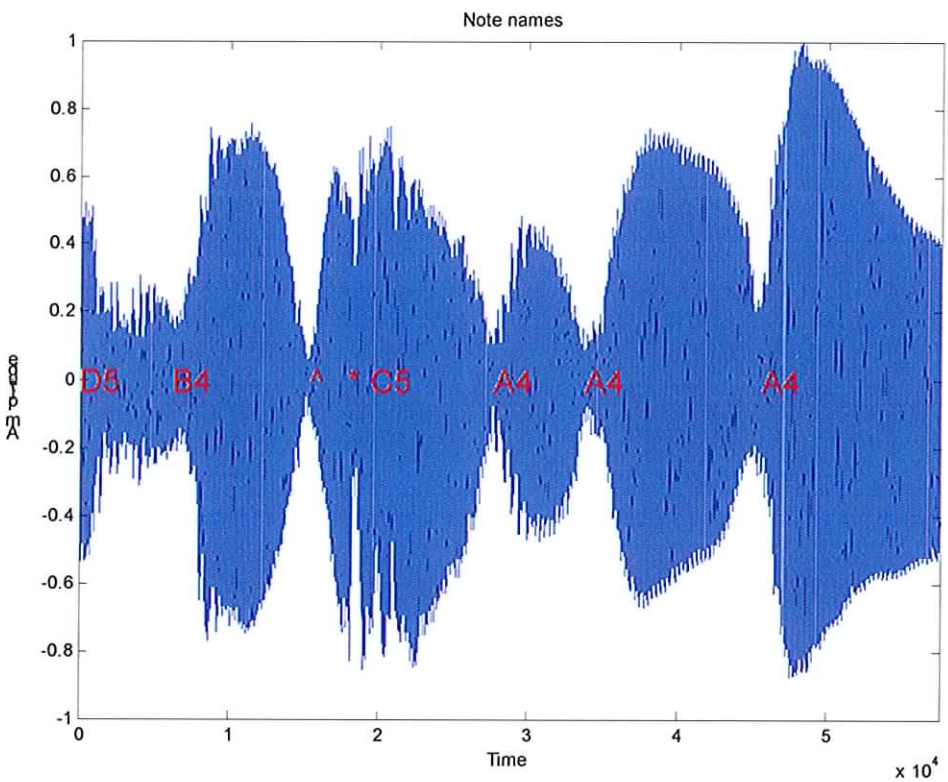

Figure 4.14: A double cut played in a fiddle tune segment

Real live recordings of tin whistle tunes from Grey Larsen's *Essential Guide to Irish Flute and Tin Whistle* [Larsen '03] were also used to test the system. Appendix A contains an index of where to find the tunes, which are listed in Table 5. The tunes are on average ten seconds long. As mentioned previously, the tin whistle is a transposing instrument. To obtain the correct score notation for a transposing instrument, the user has the option of specifying this when running the program. When this option is chosen, each detected pitch is divided by two to lower it by an octave. Otherwise the

default mode is selected and the notation will be displayed an octave too high. Percentages were calculated using Equations (4.13) and (4.14) above. Results were very good for regular note onsets with an average accuracy of 93% and also for correct pitch detections with an average accuracy of 92%. These can be seen in Table 5. Comments on each tune follow the table.

Tune	Undetected	Wrong pitch	Spurious	Detected as ornament	% Correct Pitch Detection	% Correct Onset Detection
1	0/30	0	0	0	100	100
2	1/30	2/30	2/30	0	87	90
3	0	1/29	1/29	0	93	97
4	0/30	0	0	0	100	100
5	0	0	5/31	0	84	84
6	0	0	2/30	0	93	93
7	0	0	4/30	0	87	87
8	0	0	2/33	1/33	91	91
9	0/15	0	0	0	100	100
10	2/37	1/37	2/37	0	86	89

Table 5: Onset detection results for tin whistle

1. Note onsets and pitches were detected correctly.
2. The two notes detected at the wrong pitch were D5s detected as B4 and the missing note was also D5.
3. The spurious onset occurred when there was a high jump from D4 to C5. The wrong pitch was a D5 detected a semi-tone too low as a C#5.
4. All note onsets and pitches were detected correctly.
5. Three of the spurious onsets occurred when D4s played preceding leap-wise ascending cuts which are difficult to play. One of the spurious onsets was actually a cut, which was played for longer than the set ornamentation threshold.
6. One of the spurious onsets was a cut that was played for longer than the set ornamentation threshold.

7. The four spurious onsets occurred during notes preceding large jumps.
8. The first spurious onset was a cut played for longer than the set ornamentation threshold and the spurious ornament was the middle note of a triplet.
9. All note onsets and pitches were detected correctly.
10. Both spurious onsets occurred during the first note, which was a dotted crochet, 6 times longer than a regular note. A D5 was detected a semi-tone too low as C#5. Both undetected notes had very weak onsets.

The system detected two types of ornament in tin whistle tunes, the cut and the strike, which are represented in Table 6 below as c and s respectively. Percentages were calculated using Equation (4.15) above and were a little better than in the case of the fiddle with an average accuracy of 34% for strikes and 53% for cuts.

Tune	Ornament Types	Detected	Spurious	% Correct Strike	% Correct Cut	% Correct Ornament Detection
1	10 c	6 c	0	-	60	60
2	9 c	4 c	0	-	44	44
3	9 c	5 c	0	-	56	56
4	6 s	1 s	0	17	-	17
5	8 c	5 c	0	-	63	63
6	12 c	4 c	0	-	33	33
7	15 c	13 c	0	-	87	87
8	5 c 2 s	4 c 1 s	1 s	50	80	65
9	2 c	0	0	-	0	0

Table 6: Ornamentation results for tin whistle

5. One of the undetected cuts was detected as a note, and is explained in the comments on Table 5 above.
6. One of the undetected cuts was detected as a note.

9. The undetected cut was detected as a note as it was played for longer than the set ornamentation threshold. and the spurious strike was explained in the comments on Table 5 above.

The real live recordings of flute tunes used to test the system were obtained from Grey Larsen's *Essential Guide to Irish Flute and Tin Whistle* [Larsen '03]. Appendix A contains an index of where to find the tunes, listed in Table 7. The tunes are on average ten seconds long. Percentages were calculated using Equations (4.13) and (4.14) and results in Table 7 were very good for regular onsets with an average accuracy of 93% and for correct pitch detections with an average accuracy of 92%. Comments on each tune follow the table.

Tune	Undetected	Wrong Pitch	Spurious	Detected as ornament	% Correct Pitch Detection	% Correct Onset Detection
1	0	1/38	1/38	0	97	97
2	2/29	0	0	0	93	93
3	2/25	0	0	0	92	92
4	2/30	0	0	0	93	93
5	0	0	2/28	0	93	93
6	0	0	1/27	0	96	96
7	3/31	0	0	0	90	90
8	8/33	0	0	0	77	77
9	2/40	0	0	0	95	95
10	0	2/44	0	0	95	100

Table 7: Onset detection results for flute

1. The first note was detected an octave high as F#5 with a strike before it instead of F#4. A spurious C5 was detected between a B4 and D5.
2. There were two undetected F#4s both had weak onsets.
3. Both undetected notes had weak onsets.
4. The undetected F#4s had weak onsets.

5. One of the spurious onsets was actually a cut that had been played for too long.
6. The spurious onset was a strike that had been played for longer than the set ornamentation threshold.
7. The first undetected onset was the parent note of a cut, which was detected. The second was detected as a cut and the third had a weak onset.
8. The eight undetected onsets were semi-quavers and all were detected as strikes.
9. Both of the undetected notes were picked up as strikes.
10. The two notes detected at the wrong frequency were G4s detected an octave too high at G5.

The system detected two types of ornament in flute tunes, the cut and the strike, which are represented in Table 8 below as c and s respectively. Percentages were calculated using Equation (4.15) above and there was an average accuracy of 14% for strikes and 61% for cuts.

Tune	Ornament Types	Detected	Spurious	% Correct Strikes	% Correct Cuts	% Correct Ornament Detection
1	8 c	4 c	0	-	50	50
2	8 c	5 c	0	-	63	63
3	11 c	9 c	0	-	82	82
4	8 c	5 c	0	-	63	63
5	12 c	9 c	0	-	75	75
6	7 s	1 s	1 c	14	-	14
7	6 c	3 c	1 c	-	33	33
8	6 c	3 c	8s	-	50	50
9	3 c	2 c	2 s	-	67	67
10	8 c	5 c	0	-	63	63

Table 8: Ornamentation results for flute

5. One of the undetected cuts was detected as a note.

6. An undetected strike was detected as a note
8. The spurious strikes were actually semi-quaver onsets.
10. Contained two types of multi-note ornaments that are not dealt with by this system. The two spurious strikes were note onsets.

4.4 Chapter Summary

In this section the results of a novel onset and ornament detector and music transcription system for monophonic traditional Irish music have been presented. It was tested on tunes played by three of the most popular traditional Irish instruments, the fiddle, tin whistle and flute. Results for onset detection of regular notes were very good with average accuracies of 91%, for the fiddle and 93% for both the tin whistle and flute. Pitch detection results were also very good in all cases with average accuracies of 89% for the fiddle and 92% for both the tin whistle and flute.

Results for ornamentation detection were not as good. In the case of the fiddle, 38% of cuts and 25% of double cuts were detected correctly. The average accuracy for strikes was 100% but only one strike was played in the tune database. The tin whistle fared better in the case of cuts with an average accuracy of 53%, strikes were detected correctly 34% of the time. The best results for cuts were in the case of the flute where 61% were detected correctly. Only 14% of strikes were identified accurately for this instrument. Although the results for ornamentation detection were not that good they were encouraging as a first attempt. They show that despite the fact that ornamentation detection is a very challenging task, it is possible. The most difficult aspect is setting the correct ornamentation duration threshold. In some cases where ornaments were not picked up by the detection system, they were identified as regular notes with the correct frequency.

5. Conclusions and Further Work

A thorough literature review of onset detectors and music transcription systems has been presented. During this literature review, various techniques that are currently being used to approach the problem were described. As regards onset detection, these included energy and phase based techniques. Sub-band techniques have proved popular with some opting to apply different detection techniques to the high and low frequency bands. Autocorrelation and Wavelets are examples of two techniques that have been used in music transcription, however the STFT has gained more widespread use. Each of these techniques were discussed along with their advantages and disadvantages.

An introduction to the history of traditional Irish music was also presented. This included a comprehensive explanation of the different types of ornamentation used in this style of music. A description of the well documented classical ornamentation was also given and this was used as an aid in explaining the more ambiguous traditional Irish ornaments. Each of the three instruments, the fiddle, flute and tin whistle, that were used to test the system were also given an introduction. These are among the more popular instruments used by traditional Irish musicians. The note range of each instrument was also determined.

Based on this review, a system that detects note onsets, pitches and ornaments in monophonic traditional Irish music played by the fiddle, flute and tin whistle has been implemented. The proposed system is a novel hybrid approach and combines the best elements of previous onset detectors resulting in a robust onset and pitch detector. A sub-band approach was adopted from [Klapuri '99] although some changes were made. While each frequency band in Klapuris system contained at least four notes, the proposed system allocates one frequency band to each note. The musical signal was decomposed into 29 frequency bands, one for each semi-tone in the note range G3 to B5. This note range covers the range of each instrument used to test the system. Frequency analysis was achieved using a constant Q approximation, the STFT of the signal was calculated and the resulting filter outputs combined to give a series of frequency bands whose center frequencies are logarithmically spaced. Each of these frequency bands represents a musical note.

Each of these frequency bands is analyzed independently. Firstly, the amplitude envelope is calculated followed by the first order difference function just as in [Klapuri '99]. Unlike his system, the relative difference function is not calculated. An approximation of this is obtained by moving detected onsets back three samples. Similar results are achieved but it is simpler to implement. A threshold is automatically calculated for each band by analyzing the statistics of its amplitude envelope. This is achieved by implementing the threshold approximation method in [Duxbury '02]. There is one difference, while this system analyses the amplitude envelope, Duxburys system analyzed the results from the detection function. A novel idea in the proposed system was the addition of an efficient double check for missed onsets in an STFT energy summation.

Results for onset detection were very good with average accuracies of 91% for the fiddle and 93% for the flute and tin whistle. These results would have been improved slightly if the ornament detector was excluded from the implementation as occasionally a regular note, which was played for shorter than the set ornamentation threshold was incorrectly detected as an ornament.

Note pitches were calculated by extracting segments of the waveform after each onset and carrying out frequency analysis on each excerpt. This was achieved by applying an STFT with a longer window length than was used during onset detection in order to gain better frequency resolution. Pitch detection results were very good with average accuracies of 89% in the case of the fiddle and 92% in the case of the flute and tin whistle.

Ornaments were detected by setting a duration threshold. This duration was equal for both the cut and the strike and for each individual 'note' of a double cut. This threshold was chosen after thorough analysis of the traditional Irish music database and was 11 frames or 136ms long as regular notes were generally found to be longer than this. Results for ornamentation detection were not as good as those for regular onsets. In the case of the fiddle, 38% of cuts and 25% of double cuts were detected correctly. The average accuracy for strikes was 100% but only one strike was played in the tune database. The tin whistle fared better in the case of cuts with an average accuracy of 53%, strikes were detected correctly 34% of the time. The best results for

cuts were in the case of the flute where 61% were detected correctly. Only 14% of strikes were identified accurately for this instrument.

Ornamentation detection is a very challenging task and this was reflected in the results. Although the results were not exceptional, they were encouraging as a first attempt. The main difficulty proved to be the setting of an adequate ornamentation duration threshold. In some cases where ornaments were not picked up by the detection system, they were identified as regular notes with the correct frequency. During the review on traditional Irish music, it was discovered that each individual musician plays ornamentation differently. There are only so many ways you can play a regular musical note, it is a relatively standard practice. This does not apply to ornamentation, it may be classified in books but ultimately it is up to each individual musician how they would like to play it. This leads to ornamentation being extremely difficult to detect.

Further work could involve attempting to solve the problem of ornament detection. The major problem to overcome is setting the correct ornament duration threshold. This is currently a fixed value but perhaps a method of calculating an automatic threshold could be devised. Extending the number of bands in the detection system to include additional higher octaves may contribute to a possible solution. Improved temporal resolution at note onset locations may also be beneficial as most ornaments are less than ten milliseconds long.

The ornament detection system could be developed to deal with multi-note ornaments such as the roll, another commonly used ornament, particularly in the case of the fiddle. The system could also be extended to deal with other popular traditional Irish instruments such as the accordion, guitar and banjo.

The implemented system has been shown to be effective in transcribing the regular notes of fiddle, flute and tin whistle tunes. While the results for ornamentation detection are not as good, they nevertheless represent a good starting point for future work in transcribing ornamentation in the context of Traditional Irish Music.

6. References

[Abe '97] Abe T. et al., "The IF Spectrogram: a new spectral representation," in Proceedings of the International Symposium on Simulation, Visualization and Auralization for Acoustics, Research and Education (ASVA), Tokyo, Japan, pp. 423-430, 1997.

[Allen '77] Allen J.B., Rabiner L.R., "A unified approach to short time Fourier analysis and synthesis," in Proceedings of the IEEE, Vol. 66, pp.1558-1564, 1977.

[Althoff '99] Althoff R., Kelier F., Zölzer U., "Extracting sinusoids from harmonic signals," in Proceedings of the Digital Audio Effects (DAFx) Workshop, Trondheim, Norway, 1999, Norwegian University of Science and Technology and COST, pp.97-100.

[Arfib '99] Arfib D., Delprat N., "Alteration of the vibrato of a recorded voice," in Proceedings of the International Computer Music Conference (ICMC), Bejing, China, October 1999, International Computer Music Association (ICMA), pp. 186-189.

[C.P. E. Bach 1753] Bach C.P.E., *Versuch über die wahre Art das Clavier zu spielen,* Berlin, 1753.

[Bello '00] Bello J.P., Monti G and Sandler M, "Techniques for Automatic Music Transcription," in Proceedings of the first International Symposium on Music Information Retrieval (ISMIR-01), Plymouth, Massachusetts, 2000.

[Bello '03] Bello J.P., Sandler M., "Phase-based note onset detection for musical signals," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2003.

[Bilmes '93] Bilmes J.A., "Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning and Reproducing Expressive

Timing in Percussive Rhythm," MSc Thesis, Massachusetts Institute of Technology, 1993.

[Blankertz] Blankertz B., "The Constant Q Transform," available at URL http://wwwmath1.uni-muenster.de/logik/org/staff/blankertz/constQ/constQ.html

[Blood '02] Blood Dr. B., "Ornamentation," Music Theory Online, 2002, available at URL http://www.dolmetsch.com/musictheory23.htm

[Brown '89] Brown J.C. and Puckette M.S., "Calculation of a Narrowed Autocorrelation Function," *Journal of the Acoustic Society of America (JASA)*, pp. 1597-1601, April, 1989.

[Brown '91] Brown J. and Bin Zhang, "Musical frequency tracking using the methods of conventional and narrowed autocorrelation," *Journal of the Acoustic Society of America (JASA)*, May, 1991.

[Brown '92] Brown J.C., Puckette M.S., "An Efficient Algorithm for the Calculation of a Constant Q Transform," *Journal of the Acoustic Society of America (JASA)*, pp. 2698-2701, 1992.

[Chafe '85] Chafe C., Jaffe D., "Source separation and note identification in polyphonic music," Technical Report STAN-M-34, Center for Computer Research in Music and Acoustics (CCRMA), Department of Music Stanford University, 1985.

[Chan '00] Chan K, Erjongmanee S and Hong Tay C, "Real Time Automated Transcription of Live Music into Sheet Music using Common Music Notation," Final Report, Electrical and Computer Engineering Department, Carnegie Mellon, 2000.

[Cooley '65] Cooley J.W. and Tukey J.W.,"An algorithm for the machine calculation of complex Fourier Series," *Mathematics of Computation*, Vol. 19, pp. 297-301, April, 1965.

[Cooper '94] Cooper D., Ng K. C., "A Monophonic Pitch Tracking Algorithm," Technical Report 94.15, School of Computer Studies, The University of Leeds, May, 1994.

[Corkill '91] Corkill Daniel D., "Blackboard Systems," Blackboard Technology Group, Inc., available at URL http://www.cs.virginia.edu/~acc2a/techie/notes/blkbrds.htm

[Cranitch '01] Cranitch M., *The Irish Fiddle Book*, Ossian, 2001.

[Csound] "Csound," available at URL http://www.csounds.com/

[Dempster '77] Dempster A.P., et al., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B.*, Vol. 39 no. 1, pp. 1-38, 1977.

[Dolson '86] Dolson M., "The phase vocoder: A tutorial," *Computer Music Journal*, Vol. 10, no.4, pp. 14, 1986.

[Duxbury '02] Duxbury C., Sandler M., Davies M., "A hybrid approach to musical note onset detection," in Proceedings of the 5th Int. Conference on Digital Audio Effects (DAFx), Hamburg, 2002.

[Duxbury '03a] Duxbury C., Sandler M., Davies M., "Temporal segmentation and pre-analysis for non-linear time-scaling of audio," in Proceedings of the 114th Audio Engineering Society Convention, Amsterdam, 2003.

[Duxbury '03b] Duxbury C., Bello J.P., Davies M., Sandler M., "A combined phase and amplitude based approach to onset detection for audio segmentation," in Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03), London, 2003.

[Duxbury '03c] Duxbury C., Bello J.P., Davies M., Sandler M., "Complex Domain

Onset Detection for Musical Signals," in Proceedings of the 6th International Conference on Digital Audio Effects (DAFx), London, 2003.

[Ellis '96] Ellis D.P.W., "Prediction-driven computational auditory scene analysis," PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA, June, 1996.

[Flanagan '66] Flanagan J.L., Golden R.M., "Phase Vocoder," *The Bell System Technical Journal,* Vol. 45, pp. 1493-1509, 1966.

[Fleischmann '98] Fleischmann A., Ó Súilleabháin M., & McGettrick P. ed., *Sources of Irish Traditional Music c. 1600-1855*, New York, Garland Publishing Inc, 1998.

[Gainza '04a] Gainza M., Lawlor B., Coyle E., Kelleher A., "Onset detection and music transcription for the Irish tin whistle," in Proceedings of the Irish Signals and Systems Conference, Belfast, 2004.

[Gainza '04b] Gainza M., Lawlor B., Coyle E., "Single-note ornaments transcription for the Irish tin whistle based on onset detection," in Proceedings of the 7th Int. Conference on Digital Audio Effects (DAFx-04), Naples, Italy, 2004.

[Gerstner '02] Gerstner W., "Integrate-and-fire neurons and networks," in *The Handbook of Brain Theory and Neural Networks*, Second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, 2002.

[Gold '69] Gold B. and Rabiner L., "Parallel Processing Techniques for Estimating Pitch Periods in the Time Domain," *Journal of the Acoustic Society of America (JASA)*, Vol. 46, pp. 442-448, 1969.

[Goto '00] Goto M., "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. II, pp. 757-760, 2000.

[Harris '78] Harris F.J., "On the use of windows for harmonic analysis with the discrete Fourier transform," in Proceedings of the IEEE, Vol. 66, no. 1, pp. 51-83, 1978.

[Haykin '99] Haykin S., *Neural Networks*, 2nd Edition, Prentice Hall, 1999.

[Hermansky '93] Hermansky H., Morgan N., Hirsh H.G., "Recognition of speech in additive and convolutive noise based on RASTA spectral processing," in proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Minneapolis, Minnesota, 1993.

[Howard '00] Howard D., Angus J., *Acoustics and Psychoacoustics*, Elsevier Science & Technology Books, 2nd Edition, June 2000.

[Huron '89] Huron D., "Voice denumerability in polyphonic music of homogenous timbres," *Music Perception*, Vol. 6, No. 4, pp. 361-382, Summer 1989.

[Jehan '97] Jehan T, "Musical Signal Parameter Estimation," The Centre for New Music and Audio Technologies (CNMAT), Berkeley, 1997.

[Jense '01] Jense A., la Cour-Harbo A., *Ripples in Mathematics: the Discrete Wavelet Transform*, Springer, 2001.

[Kaiser '99] Kaiser G., *A Friendly Guide To Wavelets*, Birkhauser-Boston, sixth printing, 1999.

[Kelleher '05] Kelleher A., Fitzgerald D., et al., "Onset Detection, Music Transcription and Ornament Detection for the Traditional Irish Fiddle," in Proceedings of the Audio Engineering Society Convention, Barcelona, 2005.

[Klapuri '98] Klapuri A., "Automatic Transcription of Music," MSc Thesis, Tampere University of Technology, 1998.

[Klapuri '99] Klapuri A., "Sound Onset Detection by Applying Psychoacoustic Knowledge," in Proceedings of IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), Phoenix, Arizona, 1999.

[Klapuri '00] Klapuri A., Virtanen T., Holm J.M., "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," in Proceedings of COST-G6 Conference of Digital Audio Effects (DAFx-00), Verona, Italy, 2000.

[Klapuri '01] Klapuri A., Virtanen T., Eronen A., Seppanen J., "Automatic transcription of musical recordings," in Proceedings of the Consistent & Reliable Acoustic Cues Workshop (SRAC-01), Aalborg, Denmark, September 2001.

[Knowlton '71] Knowlton Prentiss H., "Interactive Communication and Display of Keyboard Music," PhD Dissertation, University of Utah, 1971.

[Larsen '03] Larsen G., *The Essential Guide to Irish Flute and Tin Whistle*, Mel Bay Publications, 2003.

[Marchand '01] Marchand S., "An efficient pitch-tracking algorithm using a combination of Fourier transforms," in Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland, 2001.

[Marolt '02a] Marolt M., Kavcic A., Privosnik M., Divjak S., "On detecting note onsets in piano music," in Proceedings of the IEEE Electrotechnical Conference, MELECON, 2002.

[Marolt '02b] Marolt M., Kavcic A., Privosnik M., "Neural networks for note onset detection in piano music," in Proceedings of the International Computer Music Conference (ICMC), 2002.

[Martin '94] Martin J., *The Acoustics of the Recorder*, Moeck, 1994.

[Martin '96a] Martin K.D., "A Blackboard System for Automatic Transcription of Simple Polyphonic Music," Massachusetts Institute of Technology

Media Laboratory Perceptual Computing Section Technical Report No. 385, 1996.

[Martin '96b] Martin K.D., "Automatic transcription of simple polyphonic music: robust front end processing," in Proceedings of the 3rd joint meeting of the Acoustical Societies of America and Japan, December, 1996.

[Masri '96] Masri P. and Bateman A., "Improved modelling of attack transients in music analysis-resynthesis," in Proceedings of the International Computer Music Conference (ICMC), pp. 100-103, 1996.

[McCullough '87] McCullough, L.E., *The complete tin whistle tutor*, New York, Oak Publications, 1987.

[McGettrick '99] McGettrick P., *Technology and Irish Traditional Music*, Cork Companion to Irish Traditional Music, ed. Fintan Vallely Cork University Press, 1999.

[Misiti '97] Misiti, *Wavelet Toolbox Users Guide*, 1997.

[McNab '00] McNab R.J., Smith L.A., "Evaluation of a Melody Transcription System," in Proceedings of the IEEE International Conference on Multimedia and Expo, Vol. 2, 2000.

[Monti '00] Monti G., Sandler M., "Monophonic transcription with autocorrelation," in Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy, 2000.

[Moore '97] Moore B., Glasberg B., Baer T., "A Model for the Prediction of Thresholds, Loudness and Partial Loudness," Audio Engineering Society (AES), Vol. 45, 1997.

[O'Farrell 1804] O'Farrell, *O'Farrell's collection of natural Irish music for the union pipes*, London, John Gow, 1804; Complied, edited and reconstructed by Patrick Sky, Chapel Hill, North Carolina, Grassblade Music, 1995.

[Piszczalski '77] Piszczalski M. and Galler B.F., "Automatic Music Transcription," *Computer Music Journal*, pp. 24-31, 1977.

[Plumbley '02] Plumbley M.D., Abdallah S.A., Bello J.P., Davies M.E., Monti G. and Sandler M.B., "Automatic Music Transcription and Audio Source Separation," *Cybernetics and Systems*, Vol. 33, no. 6, pp. 603-627, 1 September 2002.

[Rumsey '94] Rumsey F., *Midi systems and control*, Music Technology Series, Focal Press, Second Edition, 1994.

[Scheirer '98] Schierer E., "Tempo and beat analysis of acoustic musical signals," *Acoustical society of America*, Vol. 103, no.1, pp. 588-601, January 1998.

[Schottstaedt] Schottstaedt Bill, "Common Music Notation," Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, available at URL http://ccrma-www.stanford.edu/software/cmn/

[Serra '90] Serra X., "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," PhD dissertation, Stanford University, 1990.

[Sethares '97] Sethares W.A., *Tuning, Timbre, Spectrum, Scale*, Springer 1997.

[Sillem '98] Sillem R., "Using Digital Signal Processing to Transcribe Polyphonic Music," Audio and Music Technology: The Challenge of Creative DSP (Ref. No. 1998/470), IEE Colloquium (1998).

[Slaney '98] Slaney M., "Auditory Toolbox for Matlab," available at URL http://www.interval.com/papers/1998-010/

[Smith '96] Smith L.S., "Onset-based sound segmentation," in Advances in Neural Information Processing Systems 8, Touretzky, Mozer and Haselmo (ers.), Cambridge, MA, Massachusetts Institute of Technology Press, 1996.

[Statsoft '03] Statsoft Inc. 2003, "Neural Networks," available at URL http://www.statsoftinc.com/textbook/stneunet.html.

[Smith '04] Smith III J.O., Abel J.S., "Equivalent rectangular bandwidth," Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, 2004, available at URL http://ccrma.stanford.edu/~jos/bbt/Equivalent_Rectangular_Bandwidth.html.

[Virtanen '00] Virtanen, Klapuri, "Separation of Harmonic Sound Sources Using Sinusoidal Modeling," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2000.

[WAV2MIDI] "WAV2MIDI," available at URL http://www.audiowork.com

[Wendt '96] Wendt C, and Petropulu A.P., "Pitch Determination and Speech Segmentation using the Discrete Wavelet Transform," in Proceedings of the IEEE International Symposium on Circuits and Systems, Vol. 2, pp.45-48, 1996.

[Windsor '00] Windsor W.L., Desain P., Honing H., Aarts R., Heijink H., and Timmers R., "On Time: The influence of tempo, structure and style on the timing of grace notes in skilled musical performance," *Rhythm perception and production*, pp. 217-223, Lisse: Swets & Zeitlinger, 2000.

[Wolfe] Wolfe, J., "Note names, MIDI numbers and frequencies," University of New South Wales, Sydney, Australia, available at URL http://www.phys.unsw.edu.au/~jw/notes.html

[Zölzer '97] Zölzer U., *Digital Audio Signal Processing*, 1997.

[Zwicker '99] Zwicker E., Fastl H., *Psychoacoustics – facts and models*, Springer, Berlin, 1999.

Appendix A

This appendix contains an index of the tunes that were used to test the system. It is divided into three sections, one for each instrument. The names the tunes were given in the results section are listed down the left hand side.

Fiddle

The fiddle tunes were obtained from [Cranitch '01], the page on which the notation can be found is on the right hand side.

1. Give us a drink of water 59
2. The Connachtman's Rambles 52
3. Dalaigh's Polka 69
4. The Humours of Carrigaholt 90
5. Denis Murphy's Slide 63
6. Cronin's Hornpipe 76
7. The Peeler's Jacket 83
8. The Hag's Purse 54
9. The top of Maol 67
10. The Lakes of Sligo 70
11. The Scartaglen Slide 64

Tin Whistle

The tin whistle and flute tunes were obtained from [Larsen '03] and again the page numbers are on the right hand side.

1. Study 1 308
2. Study 3 308
3. Study 4 309
4. Study 19 318
5. Study 7 310

6. Study 8 310
7. Study 9 311
8. Bantry Bay 152
9. Hardiman the Fiddler 135
10. Tuttle's Reel 28

Flute

1. The Lonesome Jig 126
2. Study 5 309
3. Study 6 310
4. Study 11 312
5. Study 17 316
6. Study 22 320
7. Lady on the island 353
8. Maids of Ardagh 356
9. The whinny hills of Leitrim 346
10. Hardiman the Fiddler 346