2007-06-01

The DiTME Project: interdisciplinary research in music technology

Eugene Coyle
Dublin Institute of Technology

Dan Barry
Dublin Institute of Technology

Mikel Gainza
Dublin Institute of Technology

David Dorran
Dublin Institute of Technology

Charles Pritchard
Dublin Institute of Technology

See next page for additional authors

Authors
Eugene Coyle, Dan Barry, Mikel Gainza, David Dorran, Charles Pritchard, John Feeley, and Derry Fitzgerald

The DiTME project

Interdisciplinary research in music technology

Eugene Coyle, Mikel Gainza, David Dorran, Charlie Pritchard, John Feeley and Derry Fitzgerald

Abstract

This paper profiles the emergence of a significant body of research in audio engineering within the Faculties of Engineering and Applied Arts at Dublin Institute of Technology. Over a period of five years the group has had significant success in completing a Strand 3 research project entitled Digital Tools for Music Education (DiTME), followed by successful follow-on projects funded through both the European Framework FP6 and Enterprise Ireland Commercialisation research schemes. The group has solved a number of challenging problems in the audio engineering field and has both published widely and patented a novel sound source separation invention.

1 Introduction: background to the DiTME project

In line with policy on research emanating from the Dublin Institute of Technology Strategic Plan 2001–2015, with encouragement to engage in creative interdisciplinary activity, a merger was formed at the turn of the millennium between the Faculty of Engineering and the Faculty of Applied Arts, via the School of Control Systems and Electrical Engineering and the Digital Media Centre (DMC). The intended aim was to bring together a cross-faculty body of researchers with interest in developing creative projects, thereby facilitating the artistic talents of staff members from the Faculty of Applied Arts with the mathematical, computing and signal processing skills of key staff members from the Faculty of Engineering.

Teaching, learning and research in music technology is a vibrant and growing discipline area, bordering upon and crossing a number of scholarly fields, including creative arts, music teaching, engineering and computing. The discipline offers exciting possibilities to school-leavers with an interest in music and technology. Rapid advancements in recent years in product development in audio and related technologies have been achieved by the application of engineering and scientific skills and know-how. As outlined in April 2000 in the report ‘Technology, foresight and the university sector’ by the CIRCA Group Europe Ltd, for the Conference Heads of Irish Universities, Digital Signal Processing (DSP) had been identified as a fast-growing and enabling core technology behind many of the recent developments in the information technology (IT) and telecommunications sectors and was noted as an area of immediate concern in respect of enhanced research growth and development at national level. Likewise, Digital Media has been recognised as one of Ireland’s strategic research and development priorities by Enterprise Ireland, Forfás, the Information Society Commission and many other independent reports.

1.1 Technological Strand 3 research application

Following an application to the Department of Education Technological Research Sector Strand 3 scheme in April 2001, the emerging audio group at DIT was successful in its application for an interdisciplinary project titled Digital Tools for Music Education (DiTME). The project proposed an integrated array of research objectives in music technology, with development of a toolkit to run on a standard multimedia PC, and a with a number of novel features which would be of benefit to both teachers and students of musicianship at all levels. These included

- a slow-down/speed-up facility which would not affect the pitch of the recorded music
- an instrument separation facility to 'comb out' a lead instrument from a piece of recorded music
- a music transcription facility to convert recorded music into music notation.

It is often beneficial for students to play along with an accompaniment whilst practising. A live accompaniment is not always available and a recording may be used instead. However, this accompaniment will have been recorded at a certain fixed tempo. *Time-scale modification* algorithms may be used to enable independent control of the playback rate (without change of key) to suit a student's current learning cycle. The desirability of such a facility for music teaching and learning had been ratified by a number of music teaching professionals in the conservatory of music at DIT.

The task of extracting individual sound sources from a number of recorded mixtures of those sound sources is often referred to as *sound source separation*. Audio source separation is a complex problem, however significant benefits and possibilities present if an audio mixture can be separated into signals that are perceptually close to the original before mixing. For example in the study of musicianship, from the most elementary stages through to virtuoso performance, the service of a competent accompanist during practice is highly desirable though not always feasible. Further, much music is scored for orchestral accompaniment but few aspiring instrumental or vocal musicians have the regular opportunity to rehearse with a professional orchestra. Music Minus One (MMO; see http://www.musicminusone.com) recognized this dilemma over 50 years ago and has recorded a library of over 400 CDs containing the most requested accompaniments (orchestral as well as piano) for a wide range of music including classical, jazz, rock 'n' roll and country and western. However, the accompaniments are recorded by professional orchestras and accompanists playing with virtuoso soloists, so the trainee musician needs to have reached a very advanced level in order to use an MMO accompaniment. If the lead instrument (or voice) could be 'combed out' of any ensemble recording then any audio CD could be transformed into an MMO format. Such a facility would be useful for both the trainee lead-part musician and the trainee accompanist.

A third highly desirable feature of the proposed music teaching and learning 'toolkit' suggested by the DIT target users is a music transcription facility. Music transcription refers to the process of converting recorded music into music notation. Existing automatic transcription systems are limited to simple monophonic (one note at a time) music. For polyphonic (more than one note at a time) the only reliable means of transcription is a very tedious manual process involving repeatedly listening to short segments of the music and comparing them to known tones. For fast music such as Irish traditional music this is often impossible. If such music can be slowed down and the lead instrument separated from the ensemble recording, then this will help to develop an automatic transcription algorithm.

2 Audio time-scale modification

Audio time-scale modification (TSM) is an audio effect that enables either speeding up or slowing down, i.e. altering the duration, of an audio signal without affecting its perceived local pitch and timbral characteristics. In other words, the duration of the original signal is increased or decreased but the perceptually important features of the original signal remain unchanged. In the case of speech, the time-scale signal sounds as if the original speaker has spoken at a quicker or slower rate. In the case of music, the time-scaled signal sounds as if the musicians have played at a different tempo. Transforming audio into an alternative time-scale is a popular and useful digital audio effect that has become a standard tool within many audio multi-processing applications.
In addition to music teaching and learning TSM has numerous applications, including:

- accelerated aural reading for the blind
- music composition
- audio data compression
- text-to-speech synthesis
- audio watermarking
- fast browsing of speech material for digital library and distance learning.

In order to achieve implementation of audio time-scale modification there are two broad categories of time-scale modification algorithms which may be applied: *time-domain* and *frequency-domain*. Time-domain techniques are computationally efficient and produce high quality results for single pitched signals such as speech and monophonic music, but do not cope well with more complex signals such as polyphonic music. Frequency-domain techniques are less computationally efficient, however they have proven to be more robust and produce high quality results for a variety of signals. A perceived drawback of frequency-domain techniques is the knowledge that they can introduce a reverberant or phasy artefact into the output signal.

In completing the research for his Ph.D. in audio time-scale modification, David Dorran focused on incorporating aspects of time-domain techniques into frequency-domain techniques in an attempt to reduce the reverberant artefact and improve upon computational demands.

2.1 Time-domain techniques

In basic terms, time-domain techniques operate by discarding or repeating suitable segments of the input waveform. This process is illustrated in Figure 1 in which a quasi-periodic waveform is time-scale compressed (reduced in duration) by discarding four periods of the original waveform. It should be appreciated that time-scale expansion could be achieved in a similar manner through repetition of short segments of the original waveform.

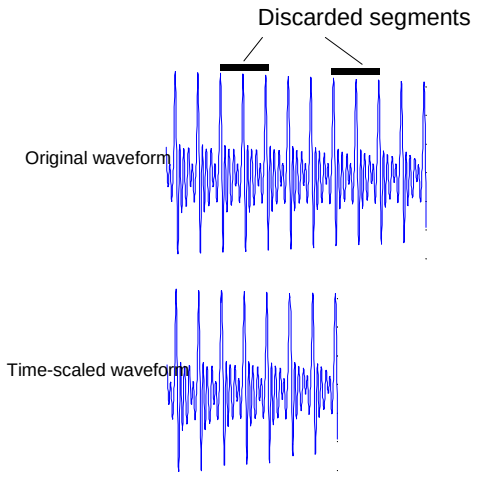

Figure 1 Time-scale compression of a quasi-periodic waveform

This example may appear somewhat trivial as it applies only to a very short sound (the original is an oboe sound of approximately 100 ms duration) that has strong periodic characteristics; however, a significant number of everyday sounds change relatively slowly over time and are therefore considered to be quasi-periodic over any 50 ms duration of the waveform. One query that often arises with regard to the periodicity of sounds is in relation to noise-like elements of a waveform, such as the ‘s’ and ‘ch’ part of the word ‘speech’ and the onset of a note of a particular instrument. It is often argued that such sounds do not contain a distinct period and therefore the discard/repeat process is not appropriate for these types of sounds; however, they can be considered periodic in the sense that the noise-like sound exists for a significant duration of time and can be viewed as the repetition of a very short noise segment over that duration. Therefore discarding/repeating short segments of these sounds will also result in time-scale expansion or compression of the sound even though they are not periodic in the strictest sense of the word.

Given the assumption of quasi-periodicity, the problem of time-scaling in the time-domain then falls into two areas: firstly, the identification of the local pitch period and secondly, identification of which segments of the original waveform to discard/repeat.

Identification of the local pitch period has received a significant amount of interest within the research community since it also forms an important part of a number

of other applications such as speaker recognition and music transcription (Kim *et al*. 2004; Plumbley *et al*. 2002). They are also used in other disciplines including biomedical signal analysis for detection of heart rate. Existing pitch period detection algorithms tend to suffer from what is referred to as 'octave errors'. For example if the pitch period was, for instance, 3 ms the algorithm may inadvertently detect a period of 6 ms, 9 ms or 12 ms, i.e. integer multiples of the actual period. However, this particular problem does not affect the quality produced by time-scaling algorithms, since the quality of the output is unaffected regardless of whether we discard one, two or three periods of the waveform. The number of periods discarded/repeated does however affect the next location for discarding/repeating the ensuing waveform segment.

The location of the discard/repeat segments is dependent principally upon the desired time-scale factor and also the duration of the segment that can be discarded/repeated. For speech, Portnoff (1981) notes that the length of each discarded/repeated segment should be longer than one pitch period (typically 4 to 20 ms) but shorter than the length of a phoneme (approximately 40 ms); these values have also been found to produce good results for music. If the duration of every segment discarded/repeated was the same, for example 10 ms, the time-scaling procedure would be very straightforward; to time-scale expand by 25 per cent, one 10 ms segment would be repeated every 40 ms; to time-scale compress by 10 per cent, one 10 ms segment would be discarded every 100 ms. In practice, since the duration of the segment being discarded or repeated must vary with the local pitch period, a slightly more complicated procedure is employed. The exact method used varies from algorithm to algorithm but all effectively keep track of the duration of the previous segment which has been discarded/repeated. If, for example, a large segment (say 16 ms) has been discarded in a particular iteration of the algorithm, then the largest segment that could be discarded in the next iteration could be forced to a time window of 4 ms, thereby ensuring that the overall time-scaling is preserved at a global level, with small variations in time-scale duration at a local level not being generally perceived to be objectionable.

The procedure outlined in the previous paragraph works well for signals that do not contain strong transient components, and is also extremely efficient in terms of computational demands. Additional care is required when transients, such as drum sounds, occur. The reason for this special treatment of transients is that, by definition, they exist for very short periods of time, i.e. less than 5 ms. If a transient segment has been discarded or repeated the result is extremely objectionable: consider the effect of removing the start of a snare drum – it would no longer sound like a snare. For this reason, time-scaling algorithms typically include a transient detection component that ensures that this problem does not arise.

2.2 Frequency-domain techniques (sinusoidal modelling and the phase vocoder)

The second technique adapted is that of sinusoidal modelling, which operates on the principle that an audio signal can be modelled by the sum of a number of quasi-sinusoidal waveforms that are slowly changing in both amplitude and frequency over time. The number of sinusoidal waveforms (or sinusoidal tracks) required to accurately represent a

particular sound depends on the type of sound being analysed. For example, the steady state portion of a flute could be well represented by only three or four tracks, whilst a timbrally rich piano would require many more. Figure 2 illustrates how an 11 ms segment of a flute waveform can be modelled by four sinusoidal tracks. Even though a single pitched example is given in the illustration, it should be appreciated that a sinusoidal model could also represent more complex sound signals.

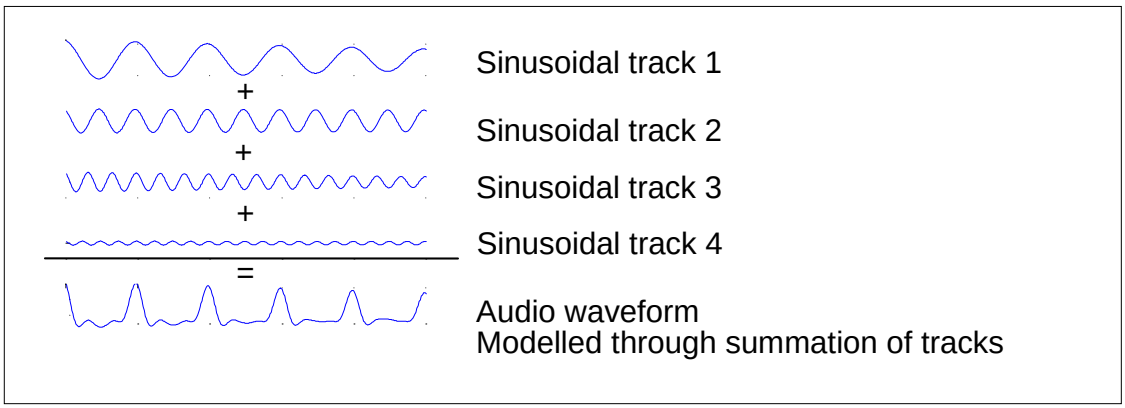

Figure 2 Modelling a flute recording by four sinusoidal tracks

The benefit of representing a complex sound through sinusoids is that these sinusoidal tracks can easily be represented as mathematical functions and can therefore be accordingly manipulated. Time-scaling via sinusoidal modelling then becomes the process of extending or compressing each individual sinusoidal track prior to summation, which could be achieved though the use of time-domain techniques described above, but is generally achieved through mathematical synthesis of sinusoidal magnitude and phase values. As the sinusoidal model is capable of representing complex multi-pitch sounds it can also be used to time-scale these types of sounds and therefore overcomes the limitations of time-domain algorithms.

The principal difficulty with sinusoidal modelling techniques is to obtain an accurate sinusoidal representation of the signal in the first place, which is a continuing area of interest within the research community. In general a reasonable representation can be obtained using a Short-Time Fourier Analysis, which can yield a perceptually accurate representation if no modifications are applied, but can however introduce objectionable artefacts when time-scaling is applied. The primary cause of these artefacts is a loss of phase coherence between sinusoidal tracks, which is perceived as a reverberant type effect in the time-scaled signal. Phase coherence is lost because of slight inaccuracies in determining the exact frequency at each instant in time of the sinusoidal tracks – these inaccuracies will always be present due to the time frequency uncertainty principle (similar to Heisenberg's uncertainty principle for mechanical systems).

Another method used which is similar to sinusoidal modelling is known as the phase vocoder. While the sinusoidal model attempts to extract a relatively small number of perceptually dominant sinusoidal tracks from a sound, the phase vocoder essentially

extracts a relatively large fixed number of sinusoids from a sound via a filterbank. The principal of extending or compressing each sinusoidal term in order to time-scale remains the same for both techniques. The advantage of the phase vocoder is that it is more robust than the sinusoidal model, since it does not require any rules to track or extract sinusoidal components. However, the filtering process employed by the phase vocoder introduces interference terms that can be problematic. The last ten years have seen a merging of the two techniques to resolve these issues (see Laroche and Dolson 1999a).

2.3 Hybrid technique

From what has been described in the previous two sections, it can be appreciated that time-domain techniques are efficient but rely on the presence of a strong periodic element with the waveform being time-scaled in order to produce high quality results; frequency-domain techniques are more robust, in that they can be applied to more general signals, but they are less computationally efficient and introduce an objectionable artefact into the time-scaled output. A hybrid approach, developed by David Dorran (2005), attempts to achieve the benefit of both time and frequency approaches to improve upon the quality of output and reduce computational demands.

The hybrid technique takes advantage of a degree of flexibility that exists in the choice of phase used during synthesis of each sinusoidal track within frequency-domain approaches. A thorough mathematical analysis shows that deviating from the mathematically ‘ideal’ phase values results in amplitude and frequency modulations entering each sinusoidal component. However, an empirical psycho-acoustic analysis (Zwicker and Fastl 1999) has shown that the human auditory system is insensitive to slight modulations in both amplitude and phase. Using these results, the maximum phase deviation (or tolerance) which can be introduced without introducing audible artefacts has been established. This phase tolerance can then be used to ‘push or pull’ the sinusoidal tracks back into a phase coherent state, thereby removing the reverberant artefact associated with frequency-domain techniques. The set of target or ‘coherent’ phases are actually taken from the original signal, since these phases are guaranteed to preserve the phase relationship between sinusoids without the introduction of reverberation. The choice of these sets of target phases is extremely important, since a ‘good’ set of target phases will reduce the transition time for sinusoidal tracks being out of phase to being back in perfect phase coherence; a shorter transition time reduces the amount of reverberation introduced. The technique used to identify the best set of target phases is based upon ‘correlation’, which is also used within time-domain techniques to identify the local pitch period.

The current implementation of the hybrid system is particularly efficient for relatively small time-scale factors. Figure 3 illustrates its computational advantage when compared to an improved phase vocoder (Laroche and Dolson 1999b) – an implementation of the phase vocoder which draws on sinusoidal modelling techniques.

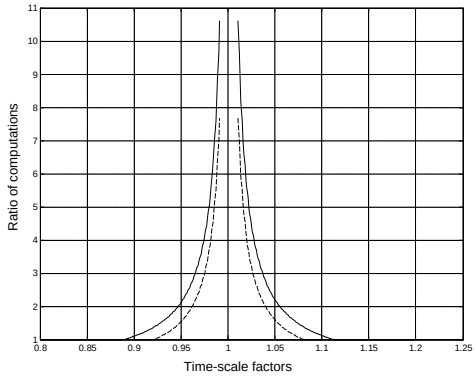

Figure 3 Ratio of computations required for the improved phase vocoder approach to the number of computations required using the hybrid approach

Subjective listening tests have also shown that the hybrid approach produces a higher quality of output to frequency-domain techniques for speech signals. No significant improvement was observed for music signals. This was attributed to the fact that music generally contains more reverberation than speech, therefore the introduction or reduction of a relatively small amount of reverberation is not objectionable. Tables 1 and 2 present the results obtained from 14 subjective listening tests. It can be seen that the algorithm is both robust and efficient and produces high quality results for both speech and a wide range of polyphonic audio. These attributes make it particularly suitable for the time-scale modification of general audio where no prior knowledge of the input signal exists, for example, during the time-scale modification of movies or television/radio adverts, in which both speech and/or music are typically present.

Test subjects indication	**Percentage of total**
Hybrid much better than phase vocoder	33.0%
Hybrid slightly better than phase vocoder	43.5%
Hybrid equal to phase vocoder	18.0%
Hybrid slightly worse than phase vocoder	5.5%
Hybrid much worse than phase vocoder	0.0%

Table 1 Summary of listening test results comparing the use of the hybrid approach against a phase vocoder approach for the time-scale modification of speech for factors in the range 0.6–1.75.

Test subjects indication	Percentage of total
Hybrid much better than phase vocoder	7.5%
Hybrid slightly better than phase vocoder	25.0%
Hybrid equal to phase vocoder	42.5%
Hybrid slightly worse than phase vocoder	20.0%
Hybrid much worse than phase vocoder	5.0%

Table 2 Summary of listening test results comparing the use of the hybrid approach against a phase vocoder approach for the time-scale modification of music for factors in the range 0.6–1.75.

3 Sound source separation

Sound source separation refers to the task of extracting individual sound sources from some number of mixtures of those sound sources. As an example, consider the task of listening in humans. We have two ears: this means that our auditory cortex receives two sound mixtures, one from each ear. Through complex neural processing, the brain is able to decompose these mixtures into perceptually separate auditory streams. A well-known phenomenon known as the 'Cocktail Party Effect' (Cherry 1953) illustrates this process in action. In the presence of many speakers, humans exhibit the ability to tend to or focus on a single speaker despite the surrounding environmental noise. In the case of music audition we exhibit the ability to identify the pitch, timbre and temporal characteristics of individual sound sources within an ensemble music recording. This ability varies greatly from person to person and can be improved with practice but is present to some degree in most people. Even young children whilst singing along to a song on the radio are carrying out some form of sound source separation in order to discern which elements of the music correspond to a singing voice and which do not.

In engineering the same problem exists. A signal is observed which is known to be a mixture of several other signals. The goal is to separate this observed signal into the individual signals of which it is comprised. This is the goal of our research. In particular, our research is concerned with separating individual musical sound sources from ensemble music recordings for the purposes of audition, analysis, and transcription. Observing only the mixture (or mixtures) of these instruments, i.e. 'the song', we aim to recover each individual sound source present in the song. The applications of source separation include the following.

- *Music education*: A common problem for amateur musicians is that of identifying exactly which instrument is playing which note or notes in polyphonic music. A sound source separation facility would allow the user to take a standard musical recording such as a song on a compact disc, and extract an individual instrument part.

- *Music transcription*: Transcription is the process of transforming some set of audio events into some form of notation. In the case of music, it involves creating a musical score from audio. This task is usually carried out by humans and is both expensive and laborious. Computerised music transcription tools do exist but are limited to monophonic transcription, and are not yet highly perfected. Sound source separation allows a polyphonic mixture to be decomposed into several monophonic mixtures thus allowing current transcription techniques to be applied.

- *Audio analysis*: In many real-world scenarios, audio recordings can often be corrupted by unwanted noise from sound sources which are proximal to the source of interest. Forensic audio analysis is one such example. Source separation can facilitate the isolation of particular sounds of interest within badly corrupted recordings.

- *Remixing and up mixing*: Multi-channel audio formats are becoming increasingly popular, such as the Dolby 5.1 and DTS surround sound formats which have become standards in the film industry and are gaining ground in the music industry too. Up mixing is the process of generating several reproduction channels out of only one or two mixtures. Old films and music, for which the multi-track recordings are unavailable, could be remastered for today's modern formats.

3.1 Existing approaches

Currently, the most prevalent approaches to this problem fall into one of two categories, Independent Component Analysis (ICA) (see Hyvarinen 1999 and Casey 2000), and Computational Auditory Scene Analysis (CASA) (see Rosenthal and Okuno 1998). ICA is a statistical source separation method which operates under the assumption that the latent sources have the property of mutual statistical independence and are non-gaussian. In addition to this, ICA assumes that there are at least as many observation mixtures as there are independent sources. Since we are concerned with musical recordings, we will have at most only two observation mixtures, the left and right channels. This makes pure ICA unsuitable for the problem where more than two sources exist. One solution to the degenerate case (where sources outnumber mixtures) is the DUET algorithm (Jourjine *et al*. 2000; Rickard *et al*. 2001). This approach assumes that latent sources are disjoint orthogonal in the time-frequency domain. This assumption holds true for speech signals but not for musical signals, since western classical music is based on harmony which implies a significant amount of time-frequency overlap. CASA methods on the other hand, attempt to decompose a sound mixture into auditory events which are then grouped

according to perceptually motivated heuristics (Bregman 1990), such as common onset and offset of harmonically related components, or frequency and amplitude co-modulation of components.

3.2 Azimuth Discrimination and Resynthesis

In the following section, we present a novel sound source separation algorithm called ADRess (Azimuth Discrimination and Resynthesis) which was developed at DIT in 2003 (Barry *et al*. 2004a and 2004b). The algorithm which requires no prior knowledge or learning, performs the task of separation based purely on the lateral displacement of a source within the stereo field; in other words, the position of the sound source between the left and right speakers. The algorithm exploits the use of the 'pan pot' as a means to achieve image localisation within stereophonic recordings. As such, only an interaural intensity difference exists between left and right channels for a single source. Gain scaling and phase cancellation techniques are used to expose frequency dependent nulls across the azimuth domain, from which source separation and resynthesis is carried out.

3.2.1 Background

Since the advent of multi-channel recording systems in the early 1960s, most musical recordings are made in such a fashion, whereby *N* sources are recorded individually, then summed and distributed across two channels using a mixing console. Image localisation, referring to the apparent position of a particular instrument in the stereo field, is achieved by using a panoramic potentiometer. This device allows a single sound source to be divided into two channels with continuously variable intensity ratios (Eargle 1969). By virtue of this, a single source may be virtually positioned at any point between the speakers. So localisation in this case is achieved by creating an interaural intensity difference (IID) – a well-known phenomenon (Rayleigh1875). The pan pot was devised to simulate IIDs by attenuating the source signal fed to one reproduction channel, causing it to be localised more in the opposite channel. This means that for any single source in such a recording, the phase of a source is coherent between left and right, and only its intensity differs. It is precisely this feature that enables us to perform separation. Figure 4 shows a typical scenario for panning multiple sources in popular music.

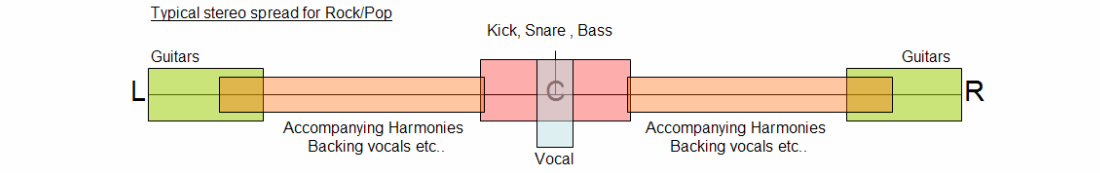

Figure 4 An example of the likely pan positions of sources in popular music

3.2.2 Method used in ADRess

A stereo recording contains two channels only (typically left and right), but any number of sources can be virtually positioned between the left and right speakers by varying the relative amplitude in each channel for a particular source. The problem is then to recover an arbitrary number of sources from only two mixtures. In order to achieve source

separation in ADRess a raised cosine window is applied to a frame of 4,096 samples of audio in each channel. A Fast Fourier Transform (FFT) is then performed, taking us into the complex frequency domain. This yields 2,048 linearly spaced discrete frequency bands of width 10.71 Hz. For each band, iterative gain-scaling is applied to one channel so that a source's intensity becomes equal in both left and right channels. A subtraction of each complex band in each channel at this point will cause that source to approach a local minimum due to phase cancellation. The cancelled source is then recovered by creating a 'frequency-azimuth' plane, which is analysed for local minima along the azimuth axis. These local minima represent points at which some gain scalar caused phase cancellation. It is observed that at some point where an instrument cancels, only the frequency components which it contained will show a local minima. The magnitude and phase of these minima are then estimated and an IFFT in conjunction with an overlap add scheme is used to resynthesise the cancelled instrument. This process is carried out on every frame of audio independently for the left and right channel for all time. Figure 5 shows this process in action for a single frequency band centred on K = 110Hz. In this example, the left channel is scaled from 1 down to 0 in discrete steps of 0.01. At each iteration, the complex value of the K^{th} scaled left channel is subtracted from the complex value in the same band in the right channel. The modulus of this operation is then taken, as shown in the plot below. At some point, this value approaches to a minimum; in this case when the gain scalar = 0.42. This signifies that a source is present at this location in stereo space. The magnitude of the component for that source is calculated as $A = K_{max} - K_{min}$. This is repeated for all bands as shown in Figure 5.

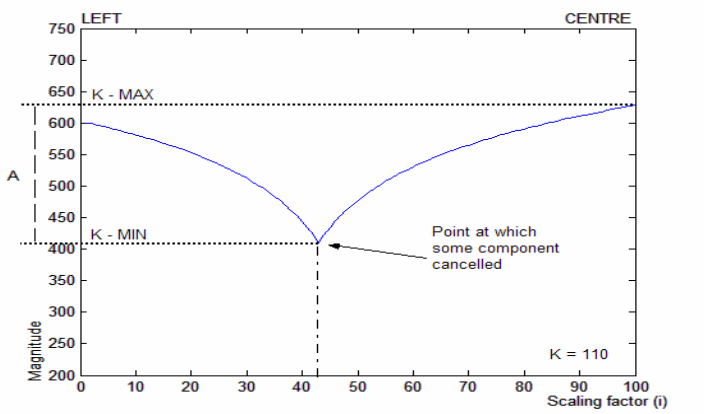

Figure 5 Gain scaling and subtraction for a single band in the frequency domain for the left side of the stereo field only. A similar operation yields the right side of the stereo field.

In order to show how frequency components belonging to a single source are clustered on the azimuth axis, two sources were synthesised, each containing five non-overlapping partials. Each source was panned to a unique location left of centre in the stereo field. Figure 6 shows the frequency azimuth plane created by ADRess to recover these sources. Frequency is depicted along the Y axis and azimuth along the X axis with amplitude represented by colour intensity.

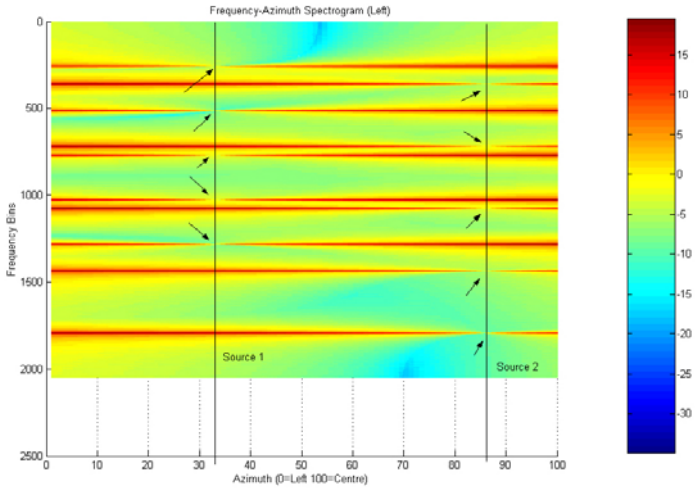

Figure 6 The frequency azimuth spectrogram shown here represents the virtual stereo space between the left channel and the virtual centre channel

It can be seen that the five frequency components from each source have their minima clustered along the azimuth axis. The frequency azimuth spectrogram shows the location of sources according to the cancellation points along the azimuth axis but, in order to resynthesise, we need the invert these nulls, since the amount of energy lost through cancellation is proportional to the actual energy contributed by the source. When the nulls are inverted we get a more intuitive representation of each individual source as demonstrated in Figure 7.

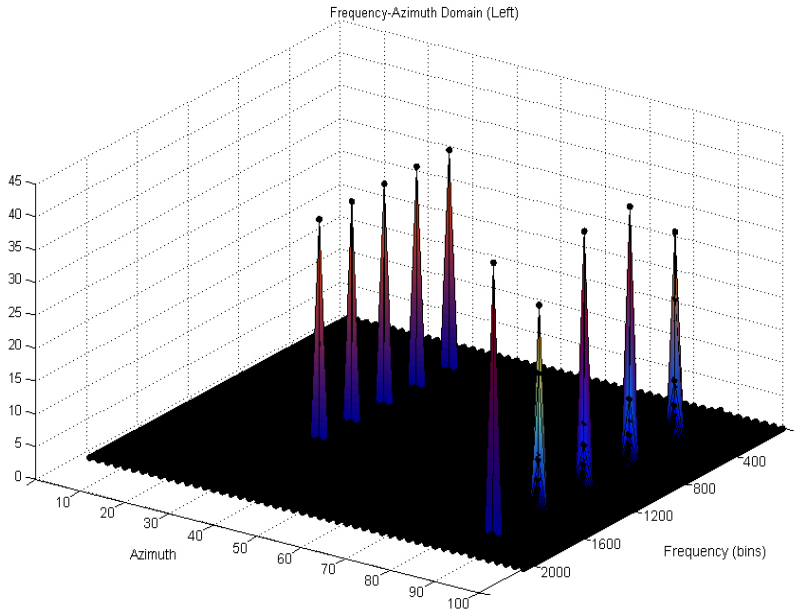

Figure 7 By inverting the nulls of the frequency azimuth composition the frequency composition of each score can be clearly seen

Figure 5 illustrates how ADRess decomposes the left channel mixture in order to reveal the frequency composition of the latent sources. It should be borne in mind that the plots in figure 6 and 7 represent the decomposition of a single frame of audio data; as each consecutive frame is processed, the composition of each source will change in both frequency and amplitude but in the majority of cases the source position (azimuth) in the stereo field will not. It is for this very reason that azimuth is used as the cue to identify each source. By summing energy at all frequencies located at different points along the azimuth axis an energy distribution plot emerges, and by doing this for all time frames a time-azimuth plot, as shown in Figure 8, is achieved. Figure 8 shows source activity in the stereo field with respect to time. A similar two dimensional visualisation updated in real time is presented to the user in order to indicate source positions in the real-time application.

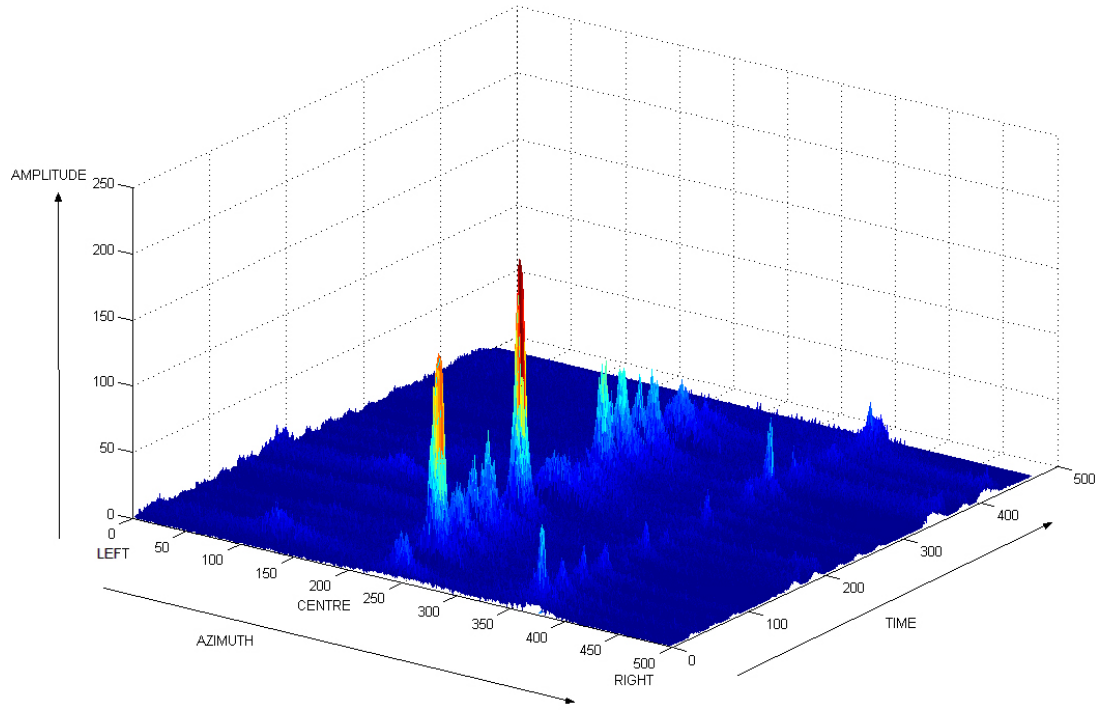

Figure 8 The plot displays the energy distribution of sources across the stereo field with respect to time. (*A source in the centre can clearly be seen as well as several other less prominent sources in the left and right regions of the stereo field*.)

The algorithm has been shown to work for a wide variety of musical recordings, some examples of which can be found at http://eleceng.dit.ie/dbarry/audio.html. The time domain plots in Figure 9 show the separation results achieved for a jazz recording containing saxophone, bass, drums and piano.

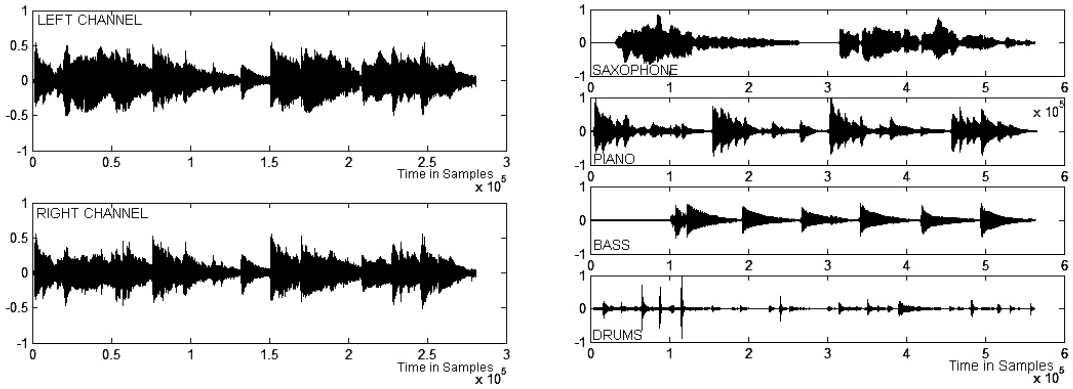

Figure 9 The two plots on the left are the left and right mixtures of a stereo recording. The four plots on the right are the individual instruments separated using the ADRess algorithm

3.3 Single channel source separation

The task of single channel source separation is significantly more difficult to achieve, nevertheless the DiTME team has given some consideration to the problem. In Barry *et al*. 2005 a method for detecting and extracting drums and other percussive signals from single channel music mixtures presented. The technique involves taking the first order log derivative of a short time Fourier transform. Following this, the number of positive tending bins are accumulated to form a percussive feature vector. The spectrogram is then modulated by this feature vector before resynthesis. Upon resynthesis only the percussive elements of the signal remain.

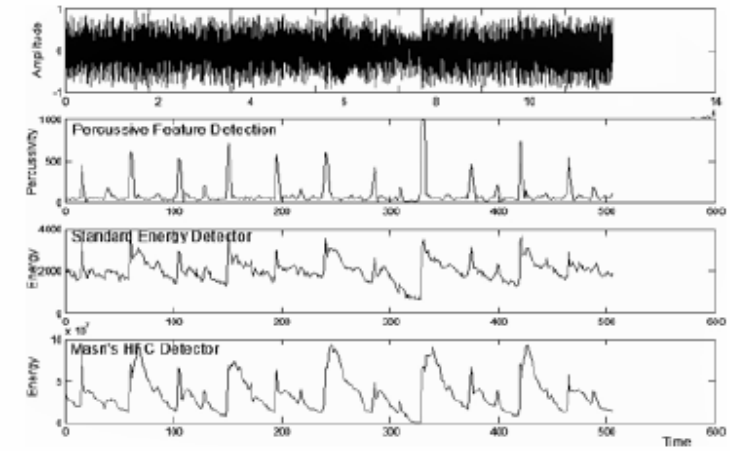

Figure 10 The first of the four plots here shows the original signal which is a piece of rock music. The second plot shows the percussive feature vector produced by our algorithm. The final two plots show the detection results of two other well-known techniques used for transient detection.

4 Music transcription within Irish traditional music

Irish traditional music has passed from generation to generation largely by oral transmission: hence the lack of transcription of this valuable cultural heritage. In researching for his Ph.D. as a member of the DiTME team, Mikel Gainza made a number of significant contributions in digital signal processing techniques to provide an understanding of the nature of audio signals in traditional music performance. Traditional music is more monophonic in nature than classical or other forms of music. It may be played as a solo performance permitting the musician to express individual nuance in style and ornamentation, or in unison with other instruments. However, simplistic harmonic accompaniment has also been incorporated in recent years. In his Ph.D. thesis 'Music Transcription within Irish Traditional Music', Gainza has identified important features of recorded notes, in particular note onset detection characteristics associated with different traditional instrument types. The 'slow' onset characteristic of the tin whistle has been carefully analysed. Ornamentation and transcription in traditional music also features in Gainza's research. In endeavouring to develop a robust automatic music transcription system, note feature characteristics must be understood. The ability to accurately detect note onset is particularly important as it provides an accurate means of recognising note commencement or event variation.

A review of existing onset detection methods in Gainza's Ph.D. (2006) concludes that the main problems encountered by existing approaches are related to frequency and amplitude modulations, in fast passages such as legato, in the detection of slow onsets, and in detecting ornamentation events. A review of existing pitch detection methods was also undertaken in this thesis, which highlights that a system that detects the different types of ornamentation within Irish traditional music has not yet been implemented. In addition, the review shows that periodicity based methods are less accurate in application to polyphonic signals.

In order to overcome the problems identified in the literature review, different applications for onset, pitch and ornamentation detection are presented in Gainza's research. These are summarised in sections 4.1 to 4.4.

4.1 Onset detection system applied to the tin whistle

First an onset detection method which focuses on the characteristics of the tin whistle within Irish traditional music was developed. This is known as the Onset Detection System applied to the Tin Whistle (ODTW). (See Gainza *et al*. 2004a.) The different blocks of the proposed onset detector are depicted in Figure 11.

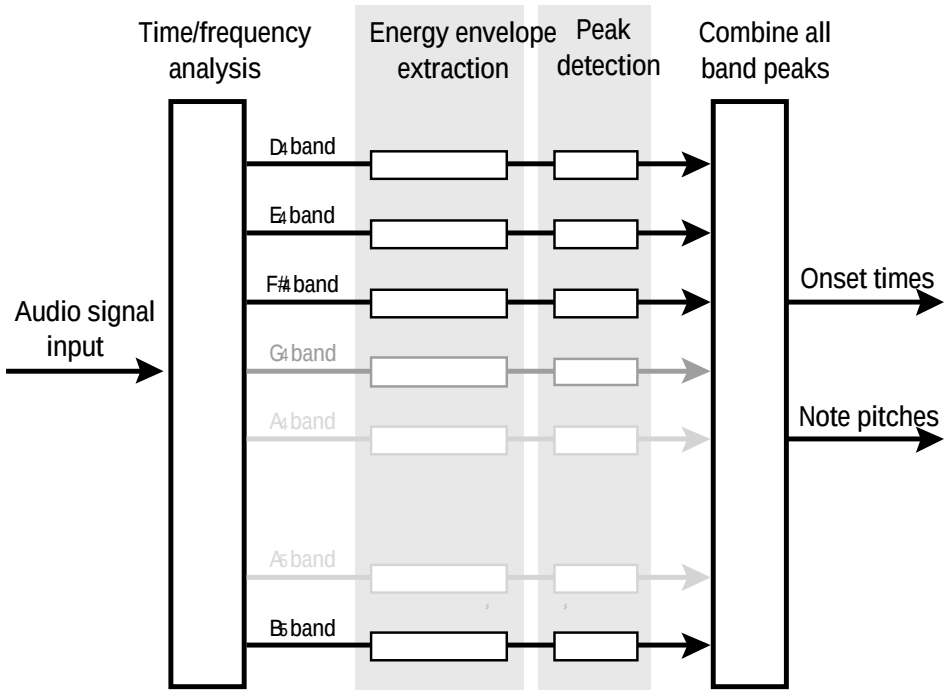

Figure 11 Overview of ODTW

A time-frequency analysis is first required, which splits the signal into different frequency bands. The energy envelope is calculated and smoothed for every band. Peaks greater than a band dependent threshold in the first derivative function of the smoothed energy envelope will be considered as onset candidates. Finally, all band peaks are combined to obtain the correct onset times.

The onset detection system utilises knowledge of the notes and modes that the tin whistle is more likely to produce, and the expected blowing pressure that a tin whistle produces per note. Problems arising in respect of legato playing in onset detection are catered for by utilising a multi-band decomposition, where one band is utilised per note. In an effort to reduce the effect of amplitude modulations, different novel thresholding methods have been implemented.

By using these methods in conjunction with an optimisation of other system parameters, the onset detection system deals with moderate signal amplitude modulations. A comparison was made of the ODTW against existing onset detection methods, configured with their respective best performing parameters: the ODTW has provided the best results.

4.2 Onset detection system based on comb filters

The ODTW provides a remarkable improvement on detecting the slow onset of the tin whistle. Nevertheless, problems of strong amplitude and frequency modulations are still present in the ODTW system. However, these limitations are overcome by a technique for detecting note onsets using FIR comb filters which have different filter delays

(Gainza *et al*. 2005). In Figure 12 a block diagram illustrating the different components of the comb filters system is depicted.

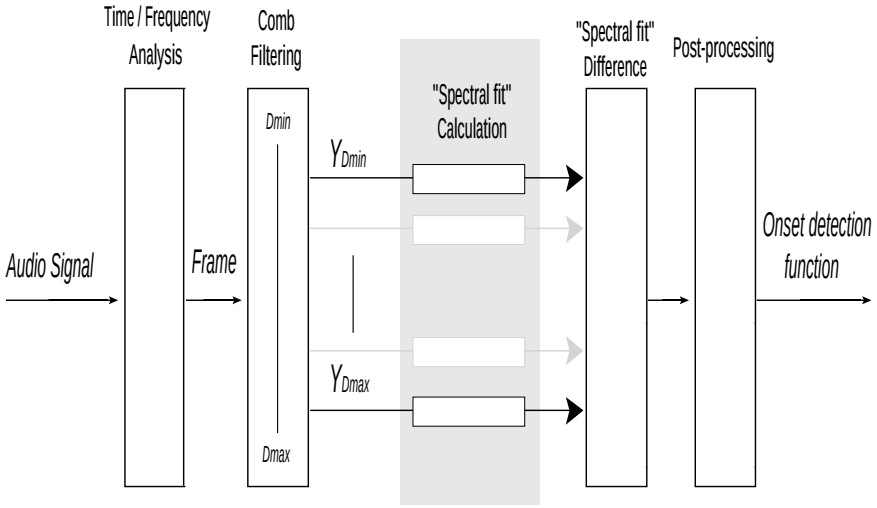

Figure 12 Onset detection system based on comb filters (ODCF)

The onset detector focuses on the harmonic characteristics of the signal, which are calculated relative to the energy of the frame. Both properties are combined by utilising FIR comb filters on a frame-by-frame basis. In order to generate an onset detection function the changes of the signal harmonicity are tracked. This produces peaks in the harmonicity changes that a new onset provides in the signal.

The method relates the harmonicity detection to the energy of the analysing frame, which is suitable for detecting slow onsets, and provides an accurate onset estimation time. The approach is robust for dealing with amplitude modulations: if the energy of the signal changes between successive frames (but not its harmonicity) the onset detection function remains stable. In addition, the method is robust to frequency modulations that gradually occur in the signal, since the signal harmonicity does not change considerably between frames.

Apart from amplitude modulations, frequency modulations can also arise in the signal, which consequently affect the onset detection accuracy. In Figure 13, the onset detection function of a tin whistle signal playing E5 is depicted in the bottom plot. The middle and top plots depict the waveform and the spectrogram of the tin whistle signal respectively, where the amplitude and frequency modulations that arise in the signal can be seen. The E5 note depicted in Figure 13 is played using a slide effect, which inflects the pitch to reach F5#, which means that a modulation between approximately 659 Hz to 740 Hz has occurred.

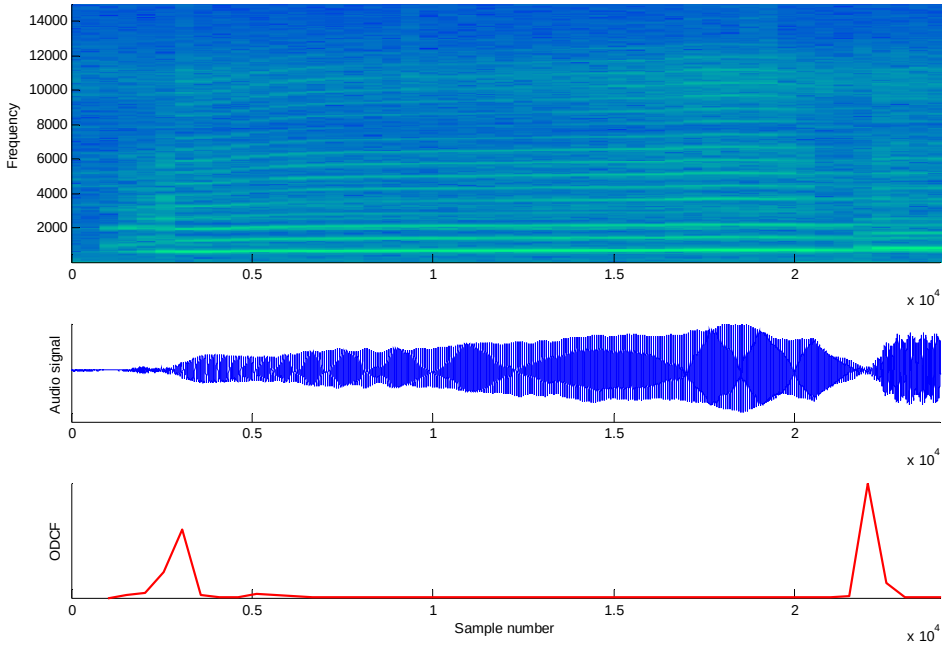

Figure 13 Onset detection function by using the ODCF (bottom plot) of a tin whistle signal (middle plot), whose spectrogram is depicted in the top plot

The onset detection function of Figure 13 depicts very distinctive peaks at the position of the onsets. It can also be seen that the slide effect does not alter the accuracy of the detection. The onset detector has been evaluated by using two different databases, which comprise tin whistle tunes and other Irish traditional music instrument tunes respectively. The results show a clear improvement upon comparison with existing onset detection approaches.

4.2 Automatic ornamentation transcription

The ODTW and ODCF systems provide a remarkable improvement on detecting the slow onsets. However, the problem related to the detection of ornamentation events in onset detection systems is not overcome by the systems, which assume that close onset candidates belong to the same onset. The latter limitation is overcome by the ornamentation detector outlined in Figure 14 (Gainza *et al* . 2004b; Gainza and Coyle 2007). The system detects audio segments by utilising an onset detector based on comb filters, which is capable of detecting very close events. In addition, a novel method to remove spurious onsets due to offset events is introduced. The system utilises musical ornamentation theory to decide whether a sequence of audio segments corresponds to an ornamentation musical structure.

The different parts of the ornamentation transcription system presented here are depicted in Figure 14. Firstly, the onset detection block is described, from which a vector of onset candidates is obtained. Next, spurious onset detections due to offset events are

removed. Following this, audio segments are formed and divided into note and ornamentation candidate segments. Next, the pitch of the audio segments is estimated. Finally, single and multi-note ornaments are transcribed.

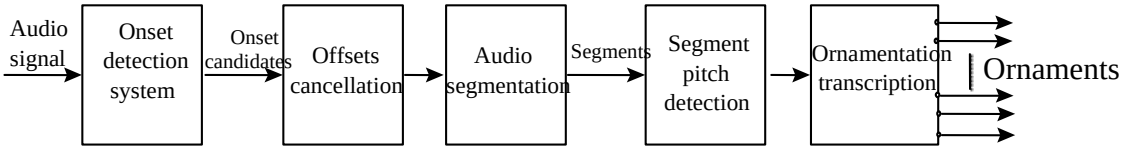

Figure 14 Ornamentation transcription system based on comb filters

Consider Figure 15 where a signal excerpt containing a roll played by a flute is depicted in the top plot. The ODF of the signal generated by utilising the ODCF is depicted in the bottom plot. It can be seen that the ODCF provides a distinctive peak at the location of the new events in the signal, which we denote as on_n.

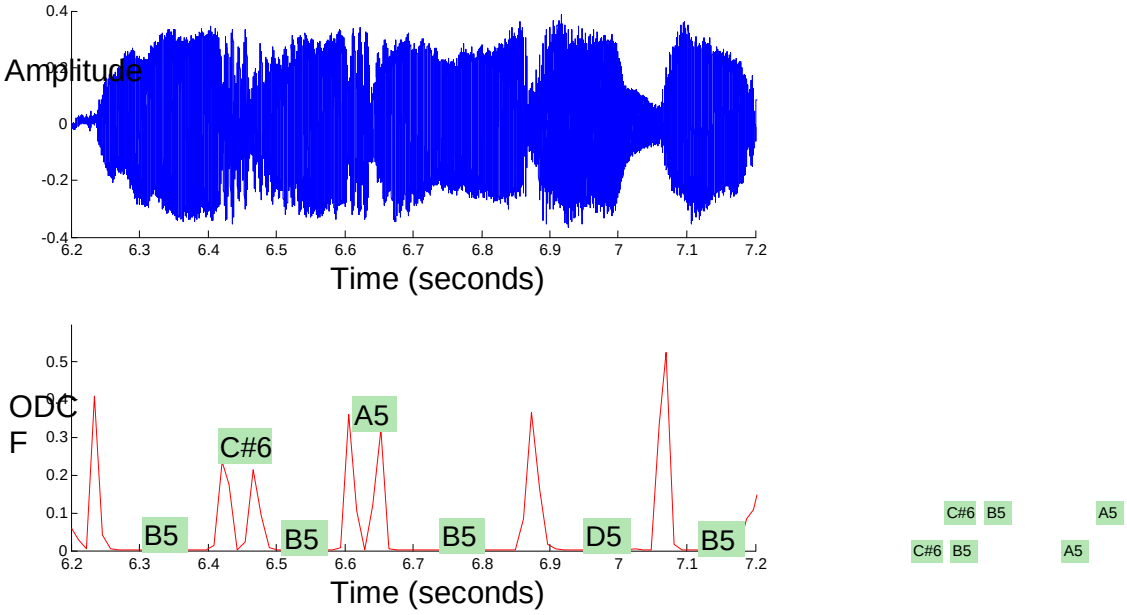

Figure 15 B5 roll – D5 – B5 sequence played by a flute

Every onset candidate on_n is matched to the next onset candidate in time order on_{n+1} to form audio segments $Sg_n = [on_n, on_{n+1}]$. Next, a table of audio segments is formed, wherein the second and third columns denote the beginning and ending of the audio segments. As an example, Table 3 shows the audio segments of the signal depicted in Figure 15.

n	***on_n (sec)***	***on_{n+1} (sec)***	***Sgn***	***P(n)***	***SNOr***	***MN Or***
1	6.235	6.42	*note*	*B5*		*Roll*
2	6.42	6.467	*orn*	*C#6*	*cut*	*Roll*
3	6.467	6.606	*note*	*B5*	*cut*	*Roll*
4	6.606	6.653	*orn*	*A5*	*str*	*Roll*

5	6.653	6.873	*note*	*B5*	*str*	*Roll*
6	6.873	7.07	*note*	*D5*		
7	7.07	…	*note*	*B5*		

Table3 Table of audio segments of Figure 15 (top plot)

Next, according to time duration, the audio segments are split into note and ornamentation segment candidates as follows:

$$\begin{aligned} Sg_n &= \text{orn} \quad \text{if } on_{n+1} - on_n < Te \\ Sg_n &= \text{note} \quad \text{if } on_{n+1} - on_n > Te \end{aligned} \qquad (1)$$

where *Te* is the longest expected ornamentation time for an experienced player, which has been analytically set to *Te= 70ms.* The Sg_n segment type *is* shown in the fourth column of the audio segments table, as can be seen in Table 3.

In order to obtain the pitch of the audio segments, a similar method to that of Brown (1992) is utilised. Following this, the fundamental frequency estimation is refined by using parabolic interpolation (Serra 1989). The pitch of each audio segment Sg_n is shown in the fifth column of Table 3, and is denoted as $P(n)$.

4.3.1 Single-note ornaments transcription (cuts and strikes)

- **The cut** momentarily increases the pitch. By considering Figure 15 for example, it can be seen that the second and third segments in Table 3 are an ornamentation and a note segment. In addition, $P(2)$= *C#6* is higher than $P(3)$ = *B5*. Consequently, B5 has been ornamented with a cut in *C#6,* and both segments together form a cut segment.
- **The strike** separates two notes of the same pitch by momentarily lowering the pitch of the second note. A strike ornament that separates two notes is also present in Figure 15 example. From Table 3 it can be derived that the fifth segment is a B5 note, which is separated from another B5 note by using the strike represented by the fourth segment.

4.3.2 Multi-note ornamentation transcription

Cranns and rolls are formed by combining ornamented and unornamented slurred notes of the same pitch.

- **The roll** is formed by a note followed by a cut segment and a strike segment. By considering Table 3, it can be seen that the combination of a B5, a cut segment and a strike segment form a roll, where the three note segments have the same pitch B5. The **short roll** version removes the first unornamented note.
- **The crann** segment structure is similar to the roll. The difference lies in the use of cuts alone to ornament the notes. The **short crann** removes the first unornamented note
- **The shake** is a four notes ornament formed by rapid alterations between the principal note and a further note one whole or one half step above it (Larsen 2003). It commences with the three ornaments and finishes with the principal

note. An example of a shake can be seen in Figure 16 (top plot), where an excerpt of a tin whistle tune is depicted. In the bottom plot the ODF generated by the ODCF is also depicted. By obtaining the pitch of those segments, a sequence of three ornaments (F#5, E5, F#5) and the principal note again E5 is obtained, which corresponds to a shake ornament.

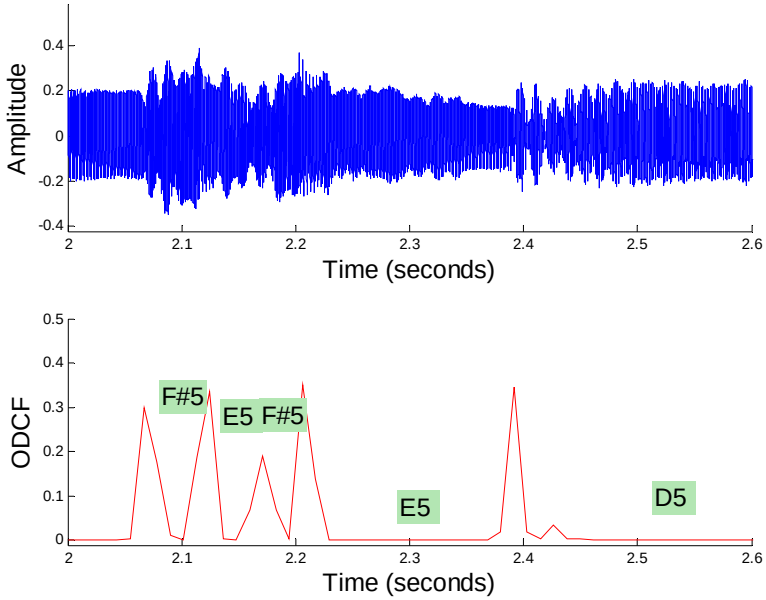

Figure 16 Example of a shake played by a tin whistle

This attempt to transcribe the most common types of ornamentation has never been previously attempted and is a particularly novel contribution to the field of onset detection and music transcription. The onset time estimation provided by this system suitably reflects Irish traditional music features, as the onset is estimated at the beginning of the ornamentation event.

Consequently, all of the difficulties encountered by existing onset detection approaches have been dealt with by the systems described in Sections 4.1 to 4.3.

4.4 Multi-pitch estimation using comb filters

When playing in unison, existing periodicity based pitch detection methods, such as FIR comb filters, might be utilised to transcribe the notes. However, with the inclusion of harmonic accompaniment the performance of these methods degrades. In an effort to detect the accompaniment chords, a multi-pitch detection system has been implemented (multi-pitch estimation using comb filters (MPECF); see Gainza *et al*. 2005b), which combines the structure of the multi-pitch detection model of Tadokoro *et al*. (2003) with the use of a more accurate comb filter and the weighting method of Martin (1982) and Morgan *et al*. (1997). The system detects the harmonic chords provided by a guitar accompaniment of a tin whistle.

In order to transcribe the musical chords played by the harmonic accompaniment, a system based on Tadokoro's model is utilised, and is depicted in Figure 17. As in Tadokoro (2003), the MPECF filter that produces an amplitude minimum represents the first detected note. Next, other notes in the audio signal are detected by iteratively connecting the output of the filter that has produced the minimum with the input of the parallel comb filter system (see Tadokoro 2003). The same filtering process is repeated again until all the notes have been extracted. After estimating the notes, an existing major or minor chord present is transcribed.

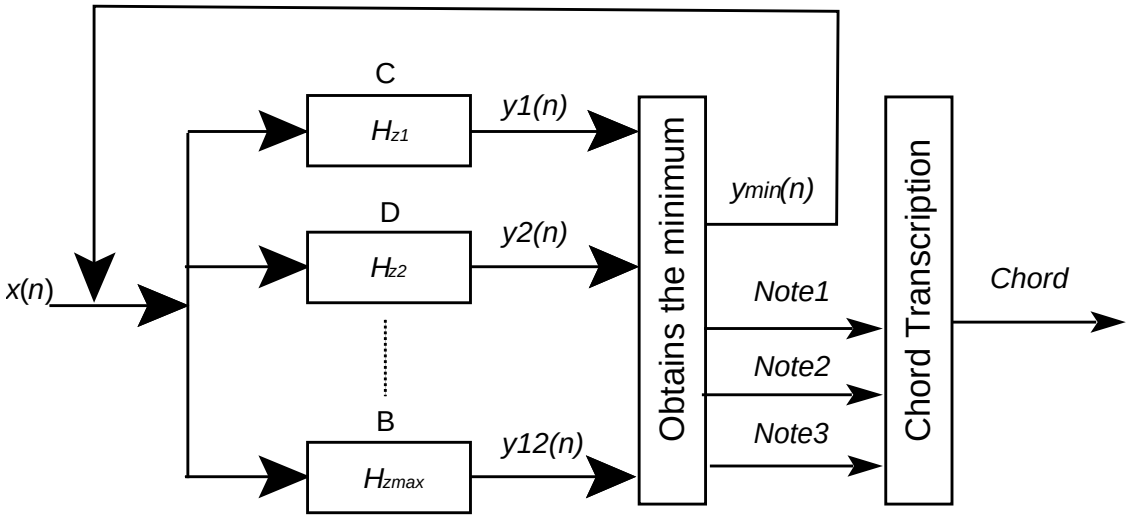

Figure 17 MPECF system for chord detection

The system has been evaluated using three different databases, comprising synthetic monophonic and polyphonic signals, real guitar chords, and mixtures of guitar chords accompanying tin whistle tunes. The results are accurate for all of the databases, where the MPECF system is capable of detecting four simultaneous notes in polyphony (three note chord and a tin whistle note).

5 New emerging Enterprise Ireland and European Framework projects

Following upon the success of the DiTME project, the Digital Audio Research group has been successful in winning research funding on a number of fronts, with success in an FP6 European Framework, an Enterprise Ireland Commercialisation funded project, and a DIT Abbest and a Strand 1 Ph.D. funded scholarship. There has been a significant increase in the number of researchers in the group, now standing at two postdoctoral fellows, one senior researcher (nearing completion of his Ph.D.), four full-time and two part-time Ph.D., and one full-time and one part-time MPhil research students.

5.1 Enriched access to sound archives through integration enrichment and retrieval

In 2005, the Centre for Digital Music at Queen Mary University London and the Audio Research Group at DIT established a consortium of seven partners including three companies and four academic institutes in a bid for a European Framework project. The bid was successful and the project officially started in May 2006 and will run until November 2008.

Many digital sound archives still have limited access facility to incumbent users. Materials are often in different formats, with related media in separate collections, and with non-standard, specialist, incomplete or even erroneous metadata. Thus, the end user is unable to discover the full value of the archived material. To expose the inherent value of the archived material, powerful multimedia mining techniques are needed, in combination with content extractors, meaningful descriptors, and visualisation tools. There is also a need to improve retrieval effectiveness. Existing retrieval systems often do not take into account the specific nature of the media content. The ability to search collections by speech or musical features is rare. Thus retrieval techniques are restricted and inflexible. To address this, multiple retrieval techniques need to be merged and deployed, and similarity and structure must be conceptualised in order to provide a usable service. An efficient and effective retrieval system needs to be grounded in semantic description, similarity, and structure in order to provide rich functionalities related to the exploration of sound archives.

Another issue is that of providing appropriate interaction with and presentation of material for the end-users. An archive used by musicians and music students, for instance, requires that the material can be manipulated or modified appropriately at playback. Archives of recorded broadcasts need to emphasise appropriate segmentation and interactive speech recognition features. In addition, the creation of tailored collections with customized material has been identified as a strong user need in access systems. These scenarios necessitate the development of enhanced and appropriate retrieval systems, as well as organisational structures and the means to interact with the presentation of materials. This demands appropriate metadata that can be automatically created in order to deliver, share or organise the archives.

This is the motivation for the two and a half year European project, ‘Enabling Access to Sound Archives through Integration, Enrichment and Retrieval’ (EASAIER). EASAIER allows archived materials to be accessed in different ways and at different levels. The system will be designed with libraries, museums, broadcast archives, and music schools and archives in mind. However, the tools may be used by anyone interested in accessing archived material – amateur or professional – regardless of the material involved. Furthermore, it enriches the access experience as well, since it enables the user to experiment with the materials in exciting new ways.

The focus of DIT’s research in this project is to provide a set of real-time access tools which will allow the user to process the retrieved audio in useful ways. DIT will

provide the following tools: time scaling, pitch shifting, source separation, noise reduction, and equalisation and enhancement tools.

5.2 Interactive Music Archive Access System

The Interactive Music Archive Access System (IMAAS) project, funded by Enterprise Ireland, is a cross-institute collaboration with the Dublin and Cork Institutes of Technology. The Irish Traditional Music Archive (ITMA) under the direction of Nicholas Carolan is also a valued partner in the project. The collaboration arose out of mutual interests and the strong links between Dr Derry Fitzgerald of CIT and the Audio Research Group at DIT. Dr Fitzgerald, who had originally obtained his Ph.D. in the field of signal processing for audio at DIT some years previously, together with Dr Matt Cranitch, consorted with the ITMA and DIT in a bid to build an interactive music archive access system. The goal is to provide remote users with a web-based access system for music archives such as the traditional music archive in Dublin. The project aims to contribute to the emerging field of music information retrieval and efficient musical descriptors. Such descriptors include time signature, key signature, tempo and tune type, to name but a few. With this information available as metadata, remote users can query large databases of music quickly and efficiently in order to retrieve only the most relevant musical data. The end user is also provided with some powerful audio processing tools to manipulate, analyse and visualise the music which has been retrieved. The project commenced in 2006 and comes to completion in November 2009.

5.3 Audio Research Group

Whilst the individual research themes of the DiTME project – audio time-scale modification, sound source separation, and music transcription (see Sections 2 to 5) – have resulted in significant contributions in advancing knowledge in their respective fields, the combined research outcomes have resulted in an even greater contribution to the field of audio research. The emerging talents of Mikel Gainza, Dan Barry, David Dorran and team mentor Eugene Coyle of DIT, and Derry Fitzgerald who completed his Ph.D. at DIT before taking up an Irish Research Council for Science, Engineering and Technology (IRCSET) scholarship under the guidance of Matt Cranitch at CIT, have provided the core of a significant research group.

With commencement of the EASAIER, IMAAS, ABBEST and VOCAL projects, the Audio Research Group at DIT currently comprises 11 researchers. In addition to the named projects, research is underway in speech synthesis, surround sound algorithms, adaptive music for gaming, musical instrument recognition, and intelligent audio environments.

In addition to a registered patent in sound source separation, to date the DIT Audio Research Group has published over 35 peer reviewed papers on various aspects of signal processing for audio. This has been supplemented by close on 20 publications by the CIT partner group. For more information about the group visit http://www.audioresearchgroup.com.

References

Barry, D., Coyle, E. and Lawlor, B. (2004a) 'Sound source separation: Azimuth Discrimination and Resynthesis', *Proceedings of 7th International Conference on Digital Audio Effects*, DAFX 04, Naples, Italy.

Barry, D., Coyle, E. and Lawlor, B. (2004b) 'Real-time sound source separation using azimuth discrimination and resynthesis', *Proceedings of 117th Audio Engineering Society Convention*, Moscone Centre, San Francisco, CA.

Barry, D., FitzGerald, D. and Coyle, E. (2005) 'Drum source separation using percussive feature detection and spectral modulation', *Proceedings IEE Irish Signals and Systems Conference*, Dublin City University, Dublin.

Bregman, A.S. (1990) *Auditory Scene Analysis*, Massachusetts: MIT Press.

Brown, J.C. (1992) 'Musical fundamental frequency tracking using a pattern recognition method', *Journal of the Acoustical Society of America*, 92 (3): 1394–1402.

Casey, M.A. (2000) 'Separation of mixed audio sources by independent subspace analysis,' Proceedings of the International Computer Music Conference, August, Berlin.

Cherry, E.C. (1953) 'Some experiments on the recognition of speech, with one and with two ears', *Journal of the Acoustical Society of America*, 25 (5): 975–979.

CIRCA Group Europe Ltd (2000) 'Technology, foresight and the university sector', Conference of Heads of Irish Universities (CHIU), April.

Dorran, D. (2005) 'Audio time-scale modification', Ph.D. thesis, Dublin Institute of Technology, Dublin, Ireland.

Eargle, J.M. (1969) 'Stereo/mono disc compatibility: a survey of the problems,' *Journal of AES*, 17 (3) (June): 276–281.

Gainza, M. (2006) ‘Music transcription within Irish traditional music’, Ph.D. thesis, Dublin Institute of Technology.

Gainza, M. and Coyle, E. (2007) ‘Automating ornamentation transcription’, IEEE International Conference on Acoustics, Speech, and Signal Processing, 15–20 April, Honolulu, Hawaii.

Gainza, M., Lawlor, B. Coyle, E. and Kelleher, A. (2004a) ‘Onset detection and music transcription for the Irish tin whistle’, *Irish Signals and Systems Conference* (ISSC), Belfast.

Gainza, M., Lawlor, B. and Coyle, E. (2004b) ‘Single-note ornaments transcription for the Irish tin whistle based on onset detection’, *Proceedings of 7th International Conference on Digital Audio Effects*, DAFX 04, Naples, Italy.

Gainza, M., Lawlor, B. and Coyle, E. (2005a) ‘Onset detection using comb filters’, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York: New Paltz.

Gainza, M., Lawlor, B. and Coyle, E. (2005b) ‘Multi pitch estimation by using modified IIR comb filters’, 47th International Symposium focused on Multimedia Systems and Applications (ELMAR), Zadar, Croatia.

Hyvarinen, A. (1999) ‘Survey on independent component analysis,’ *Neural Computing Surveys*, (2) : 94–128; available online at http://www.icsi.berkeley.edu/~jagota/NCS (accessed February 2007).

Jourjine, A., Rickard, S. and Yilmaz, O. (2000) ‘Blind separation of disjoint orthoganal signals: demixing n sources from two mixtures,’ *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, June, Istanbul, Turkey

Kim, S., Eriksson, T., Kang, H.G. and Dae Hee Youn (2004) ‘A pitch synchronous feature extraction method for speaker recogition’, IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Quebec, Canada.

Laroche, J. and Dolson, M. (1999a) ‘Improved phase vocoder’, *Speech and Audio Processing*, IEEE Transactions on Speech and Audio Processing, 7 (3) (May): 323–332.

Laroche, J. and Dolson, M. (1999b) ‘New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects’, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October, New York: New Paltz, pp. 91–94.

Larsen, G. (2003) *The Essential Guide to Irish Flute and Tin Whistle*, London: Mel Bay Publications.

Martin, P. (1982) ‘Comparison of pitch detection by cepstrum and spectral comb analysis’, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Paris (7): 180–183.

Morgan, D.P., George, E.B., Lee, L.T. and Kay, S.M. (1997) ‘Cochannel speaker separation by harmonic enhancement and suppression’, *IEEE Transactions on Speech and Audio Processing*, 5 (5): 407–424.

Plumbley, M.D., Abdallah, S.A., Bello, J.P., Davies, M.E., Monti, G. and Sandler. M.B. (2002) ‘Automatic music transcription and audio source separation’, *Cybernetics and Systems*, 33 (6): 603–627.

Portnoff, M.R. (1981) ‘Time-scale modifications of speech based on short-time Fourier analysis’, *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-29 (3) (June): 373–390.

Rayleigh, L. (1875) ‘On our perception on the direction of a source of sound’, *Proceedings of the Musical Association*, Royal Musical Association, Oxford :Oxford University Press, pp. 75–84.

Rickard, S., Balan, R. and Rosca. J. (2001) ‘Real-time time-frequency based blind source separation,’ *Proceedings of ICA 2001 Conference*, 9–13 December, San Diego, CA.

Rosenthal, D.F. and Okuno, H.G. (1998) *Computational Auditory Scene Analysis*, Mahwah, NJ: LEA Publishers.

Serra, X. (1989) ‘A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition’, Ph.D. thesis, Stanford University, USA.

Tadokoro, Y., Morita, T. and Yamaguchi, M. (2003) ‘Pitch detection of musical sounds noticing minimum output of parallel connected comb filters’, Conference on Convergent Technologies for Asia-Pacific Region (TENCON), Bangalore, India.

Zwicker, E. and Fastl, H. (1999) *Psychoacoustics: Facts and Models*, 2nd edn, New York: Springer.