EXTRACTION OF EXPRESSIVE PERFORMANCE PARAMETERS FROM ACOUSTIC RECORDINGS OF PIANO MUSIC

A Earis Department of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, U.K. e-mail: earisa@cs.man.ac.uk

BMG Cheetham Department of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, U.K.

ABSTRACT

Measurable features of expressive musical performance include timing, dynamics, articulation and pedalling. This paper concerns the measurement of expressive timing and dynamics in acoustic recordings of piano music with reference to a digitised musical score (in MIDI format) of the work being performed. Synchronisation of score and recording is achieved using the continuous wavelet transform (CWT). The acoustic characteristics of individual piano notes, and their 'dual decay', are discussed. A wavelet, based on a simple acoustical model of the initial decay (the 'prompt' sound) of a piano note, is proposed. This consists of a complex wave enveloped by a linear rise and an exponential decay. The wavelet used to locate individual notes in time (expressive timing), and measure the volume (expressive dynamics) in extracts of acoustic recordings of computer synthesized (MIDI) and real performances of piano music.

1 INTRODUCTION

A skilled musical performance has technical and expressive elements. The technical elements are the playing of a musical instrument and the accurate reading of the score. The expressive elements are intentional variations in the way the notes are played, chosen by the performer to influence the outcome for the listener. Effective expressive performance involves very fine and subtle deviations from the notated score.

Measurable features of expressive musical performance include timing (note onset and offset times), dynamics (variations in the intensity of notes and chords), tone quality (timbre), articulation (the degree to which a performer detaches individual notes from one another), vibrato, pitch and, for a piano, pedalling. The nature and number of these features depends on the instruments being played and their characteristics. For a piano, the performer controls mainly timing, dynamics, articulation and pedalling.

Digital signal processing (DSP) techniques may be used to resolve musical sounds in both time and frequency. It is thus possible to analyse acoustic recordings to automatically extract expressive performance information, allowing the different expressive features to be measured and parameterised. In order to extract the expressive performance parameters, the individual notes and chords in the acoustic recording must be located in time. An algorithm is described which will automatically extract the expressive timing information with reference to a digitised version of the musical score of the work being performed.

2 RESEARCH HISTORY

Investigations in music performance date back as far as 1895[1], with much musicological research taking place in the USA in the 1930s[2]. Much has been learned about performance characterisation by studying recordings made on MIDI (Musical Instrument Digital Interface) instruments[3]. However, the information thus recorded is limited to note onset, note offset and volume. There have been a number of studies of acoustic recordings, including the analysis of timing patterns[4,5] and dynamics[6]. Acoustic recordings, which contain a much more complete range of expressive performance features than MIDI, are universally available providing a vast supply of invaluable musical information. Recordings spanning over a century are available, thus enabling the analysis of a wide range of performance periods and styles. Such studies are generally limited to piano performances. In the papers cited, the measurements were made by manually studying the waveforms of the sound recording and locating note onsets using both visual and auditory clues.

Computational analysis of musical performance has had a relatively brief research history. Research in this field falls into a number of categories[7]

- automatic transcription – the development of algorithms to synthesise a musical score from an acoustic recording.
- music information retrieval – the cataloguing and classification of musical information, for example, on the internet or for library databases.
- audio beat tracking – the development of algorithms to ‘tap’ along to the beat.
- automatic accompaniment systems.
- computerized analysis of expressive musical performance.

The earliest algorithm which attempted to extract expressive timings (both note onset and offset) and dynamics was developed by Scheirer[8]. Onset extraction for individually sounding notes was achieved by measurement of the point of peak derivative (slope) in the high frequency energy above 4kHz, assumed to be the noise of the hammer strike, and in the overall RMS power. If an onset was not detected, a comb filter, centred on the fundamental frequency and harmonics of the note, was used, and the point of peak derivative again measured. The onset was defined as being the positive-going zero crossing in the derivative, measured by sliding back in time. For multiple simultaneous-sounding tones the latter approach is taken, using a multiple bandpass filter with pass regions centred on the frequencies of the non-overlapping harmonics of the particular note being sought. Note offset time was defined to be the time at which the output from the multiple bandpass filter either falls below 10% of the peak power, or begins rising to another peak. The amplitude was defined as the peak power point in the filtered signal.

The system was evaluated using a Yamaha Disclavier MIDI piano. Scheirer concluded that the system was accurate enough to measure certain limited aspects of music performance, including tempo estimation, but not accurate enough to measure more subtle aspects of melodic-harmonic timing.

More recently Dixon[9] has attempted to extract expressive timing by using a bank of filters centred on the frequencies of the fundamental and harmonics. For each harmonic, the time of sharpest attack before the peak is found, and the note onset time is defined as the mean of the individual onset times from each harmonic. This work is still in an early stage of development.

3 THEORY AND METHOD

In order to characterise an acoustic recording, individual notes and chords must be identified and related to the musical score. The score is digitised in MIDI format, and from this is extracted a list of events (notes or chords) and their notated duration, to be located in the acoustic waveform. Spectral analysis techniques can be performed on the acoustic waveform in order to determine the frequencies present and how these progress in time. These frequencies are then correlated with the expected frequencies of the events as predicted from the digitised score. The nature of the musical signal leads to a number of constraints on the time-frequency resolution required.

(i) The frequency resolution required is determined by the interval between adjacent semitones which is proportional to the frequency of the lower semitone. The spacing between the fundamental frequencies of individual semitones varies from about 1.6Hz at the bottom of the piano keyboard to over 120Hz at the top with the fundamental frequencies ranging from 27.5Hz to just over 4kHz.

(ii) The shortest note duration in rapid note passages is around 50ms. Expressive timing variations can be very subtle, the human ear being capable of perceiving differences of as little as 10ms [9], depending on the speed of the music, or frequency of note onset.

One approach to the spectral analysis could be the short time Fourier transform (STFT) with a fixed time window size, the size of which determines the time and frequency resolution. The STFT at a particular point in the time-frequency plane is defined as follows

$$STFT(\tau, f) = \int_{-\infty}^{\infty} x(t) g(t-\tau) e^{-j2\pi ft} dt \qquad (1)$$

where $x(t)$ is the signal, $g(t)$ is a window function centred on $t = 0$, f is the frequency and τ is the location in time.

In practice the Fourier transform would be implemented digitally using the discrete Fourier transform (DFT) or fast Fourier transform (FFT). The STFT can be viewed as the passing of the signal $x(t)$ through a series of equally-spaced bandpass filters. The fixed window (implying a fixed bandwidth) does not give the required time-frequency resolution over the complete frequency range of the piano.

The approach presented in this paper is to replace the STFT by a discretised continuous wavelet transform (CWT)[10]. The continuous wavelet transform at a particular point in the time-frequency plane is defined as follows

$$C(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) w^{*}\left[\frac{t-b}{a}\right] dt \qquad (2)$$

where $w(t)$ is a complex-valued wavelet function. Frequency is indicated by the scaling factor a and the location in time by b. The scaling factor is inversely proportional to frequency.

The wavelet function is a scaled (stretched and squeezed) and shifted (moved) version of a 'mother wavelet'. A commonly used mother wavelet is the Morlet wavelet.

$$w(t) = \frac{1}{\pi^{1/4}} e^{j2\pi f_0 t} e^{-t^2/2} \qquad (3)$$

where f_0 is the central frequency of the wavelet. This equation defines a complex wave within a Gaussian envelope, and satisfies the following equation.

$$\int_{-\infty}^{\infty} |w(t)|^2 dt = 1 \qquad (4)$$

For a fixed value of a, $C(a,b)$ is the convolution between $x(t)$ and $\left(1/\sqrt{a}\right)w(-t/a)$ evaluated at time $t = b$. In practice, the signal $x(t)$ is sampled at frequency f_s and the wavelet function may also be sampled at f_s so that a discretised version of $C(a,b)$ may be computed as the discrete time convolution between the signal and discretised wavelet function. This can be implemented by passing the signal through a bandpass digital filter with impulse response $\left(1/\sqrt{a}\right)w(-t/a)$.

The shape of the Morlet wavelet described in equation (3) is not optimized to the particular nature of piano notes.

A real piano note is illustrated in Figure 1.

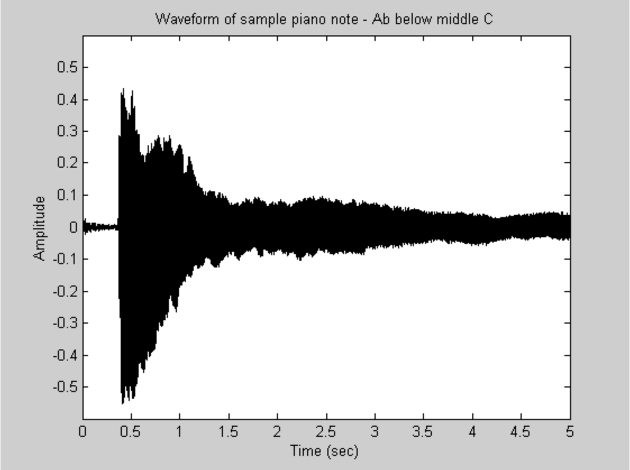


Figure 1. A real piano note
(A flat below middle C)

There have been a number of studies in the acoustical properties of individual piano notes[11,12]. When a piano key is depressed, the hammer hits the string, and this has the effect of exciting combinations of the modes of vibration of the string. In an ideal string, these modes would vibrate at frequencies that are integer multiples of the fundamental frequency of the string. However, in real strings, this relationship is not exactly harmonic, due to a number of factors including the stiffness of the string and imperfections in the string. For a piano, the higher modes become progressively sharper, and they are known as partials.

After the hammer and string contact has ended, the amplitudes of the excited modes, or partials, decay as time goes by. The tone quality, or timbre, of the note is determined by the relative frequencies and amplitudes of the partials, and their evolution in time[13].

The piano note consists of three main parts:

(i) The onset transient. This includes the sound of the hammer hitting the string as well as other noise from the instrument keys and case. The hammer and string remain in contact for as much as 4ms.

(ii) The waveform of the sound after the hammer is released displays a ‘dual decay’. Initially, the sound decays quickly. This is known as the ‘prompt sound’. This is followed by a slower decay known as the ‘aftersound’. Whilst both of these decays are exponential in nature, there is an identifiable change from the prompt decay to the aftersound decay. There are two factors that are believed to contribute to this process[14].

a) The polarization of the vibrations of the strings. Initially, the dominant mode of vibration of the strings is vertical (in the direction of the hammer hit) although there is some initial horizontal motion, thought to be due to irregularities in the hammer and string. Energy can transfer between the two polarizations, and ultimately, the horizontal motion becomes dominant[13].

b) The multiple stringing of individual piano notes. When the strings are initially struck, the strings vibrate in phase, thus creating a strong driving force at the bridge which means that energy is lost more quickly. Soon afterwards, when the strings are not all exactly in phase, the corresponding bridge driving force is less, hence the rate of dissipation of energy is less and the sound decays more slowly[15].

The time after onset at which the decay rate changes for a particular note is expected to be the same, irrespective of the initial amplitude i.e. it is believed that the duration of the prompt sound does not depend on the amount of initial displacement of the string[15].

(iii) When the key is released, the damper falls on the string and resonance ceases.

The decay of the first four harmonics of the note given in Figure 1 above are illustrated below.

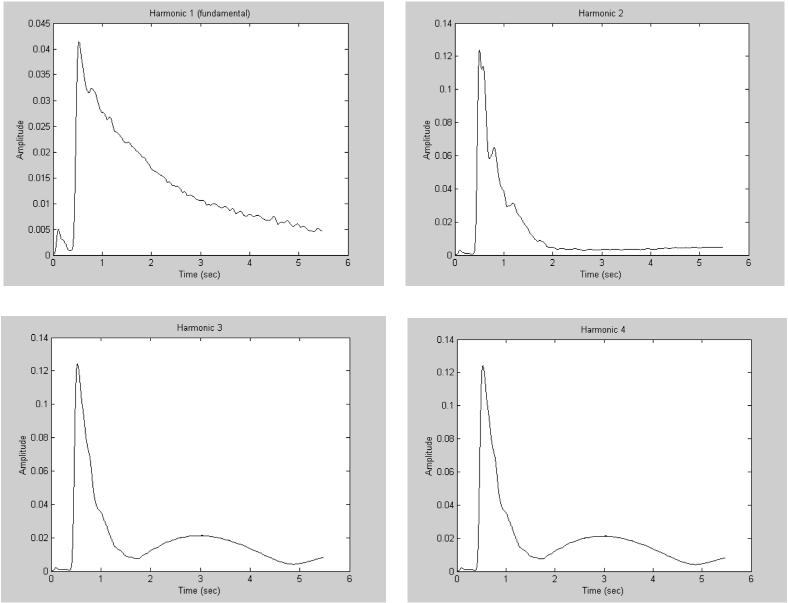

Harmonic 1 (fundamental)
Harmonic 2
Harmonic 3
Harmonic 4

Figure 2. Envelope of first four harmonics of a real piano note
(A flat below middle C)

These graphs illustrate the complexity of the evolution of the constituent harmonics of individual piano notes. However, a common feature is that the first part of the decay (the ‘prompt’ sound) is generally exponential in nature. It is this feature that is used in the construction of a new wavelet that more closely resembles the properties of a harmonic of an individual piano note. The wavelet consists of a complex wave enveloped by a linear rise followed by an exponential decay.

$$w(t) = \begin{cases} 0 & : \quad \mathrm{t} < t_R \\ c_w e^{j2\pi f_0 t}\left(1 - \dfrac{t}{t_R}\right) & : \quad t_R < t < 0 \\ c_w e^{j2\pi f_0 t} e^{-t/t_F} & : \quad t > 0 \end{cases} \qquad (5)$$

Where the constant c_W is calculated such that it satisfies equation (4). A mother wavelet with these features is illustrated in Figure 3.

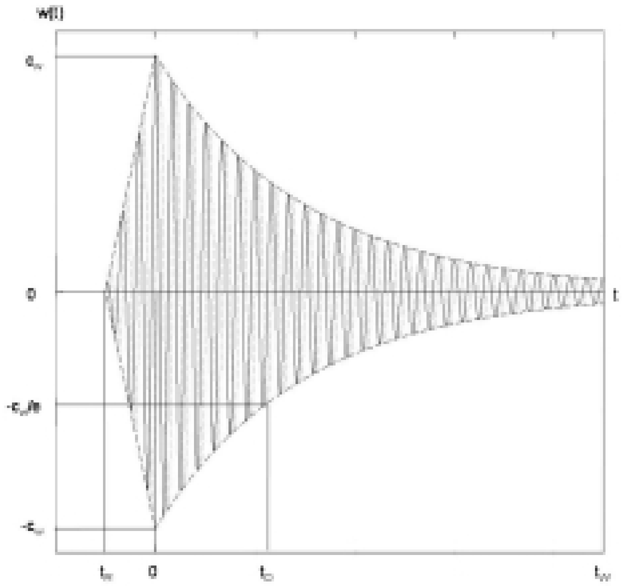


Figure 3. The wavelet with envelope characteristic of piano tone
(solid line – real part, dotted line – imaginary part, dashed line – envelope)

The wavelet function is sampled at f_S and windowed to a convenient length. The number of cycles in the rise is

$$k_R = -t_R f_0 \tag{6}$$

and the number of cycles in the exponential decay from t = 0 to $t = t_D$, the time with which the wavelet decays by a factor of $1/e$ from its maximum amplitude, is

$$k_D = t_D f_0 \tag{7}$$

Individual piano notes, covering the complete range of the piano keyboard, were recorded for a number of acoustic grand pianos. A linear phase FIR bandpass digital filter was used to extract each of the first five harmonics for each note. The waveform of each individual harmonic was rectified and low pass filtered to produce the envelope. It is generally accepted that the time interval from the point of maximum amplitude of the envelope to the point at which the amplitude has decayed by 20dB contains solely the 'prompt' sound[16]. A straight line was fitted to the log of the envelope over this time interval. The gradient of the straight line was measured and the value of t_D calculated. To measure the rise time parameter, a straight line was fitted to the envelope of the rise from 10% to 90% of the maximum amplitude. The point of maximum amplitude of the harmonic was defined as the point in time at which the fitted rise and decay lines cross and hence t_R was calculated.

For each harmonic of each note, t_R and t_D were measured, and k_R and k_D were calculated. The mean values of k_R and k_D were 7.9 (standard deviation 7.2) and 46.0 respectively (standard deviation 21.8). These mean values were chosen and the parameters of the wavelet.

An algorithm was developed to extract the expressive timing and dynamics information from an acoustic recording, using the discretised CWT with the wavelet function described above. The range in time over which the CWTs are evaluated in order to detect a particular event is calculated from a tempo estimation based on a window of previous measured timings.

For a particular note, individual CWTs were computed for each of the first five harmonics of the note being detected. The envelope of each CWT is obtained by rectifying and low pass filtering. The location in time of the point of peak amplitude in the envelope gives the onset time for the harmonic. The onset time of the event is the average of the onset times of the individual harmonics.

Rather than attempting to measure the expressive dynamics of individual notes, for example in a chord, the expressive information extracted is based on event dynamics as discussed by Repp [6]. Repp argues that it may be possible to ignore the fact that most musical 'events' consist of several musical tones, and to consider that the total sound energy may be the most important factor. Event dynamics are measured by summing the squares of the amplitudes of the peak of the CWTs.

4 RESULTS

The algorithm was evaluated by measuring expressive timing and dynamics in the opening bars of Bach's *Fugue in C major* BWV 846b (from the 'Forty Eight' Preludes and Fugues) as illustrated in Figure 4.

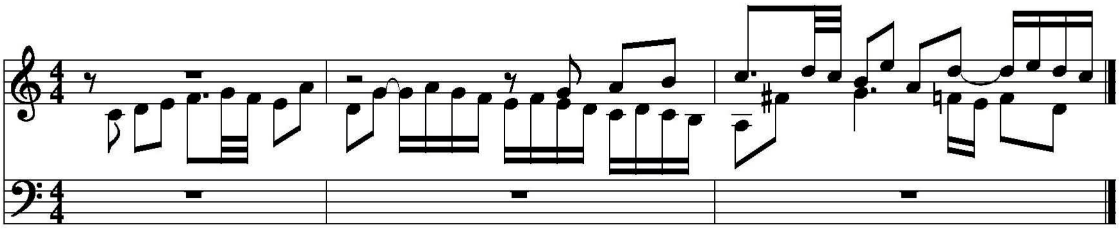
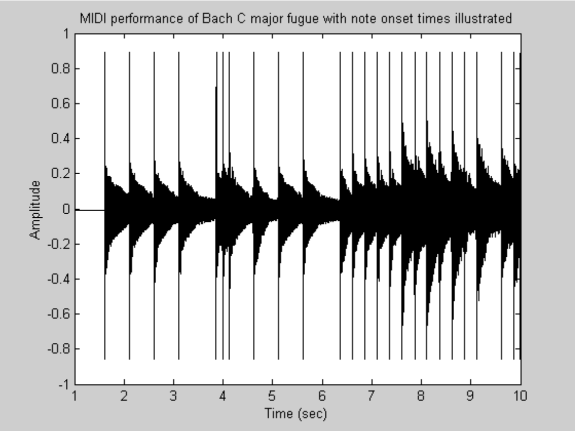
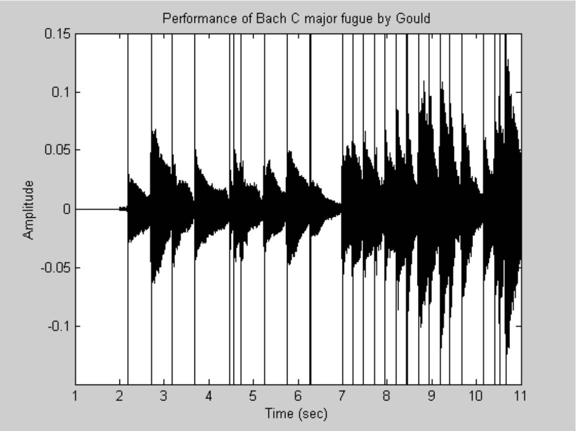


Figure 4. The opening bars of Bach's *Fugue in C major*, BWV 846b

Three different performances were evaluated
- (i) A computer-synthesised performance from a MIDI file with piano synthesizer
- (ii) A performance from an acoustic recording by Glenn Gould
- (iii) A performance from an acoustic recording by Friedrich Gulda

Graphs showing the waveforms of all three performances, together with superimposed vertical lines illustrating the algorithm measured note onset times, are shown in Figure 5.





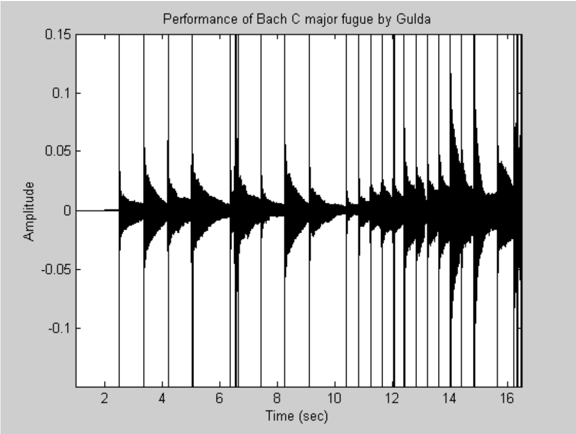


Figure 5. The waveform of the opening bars of three performances of Bach's Fugue in C major, BWV 846b, with algorithm-measured note onset times illustrated by vertical lines.

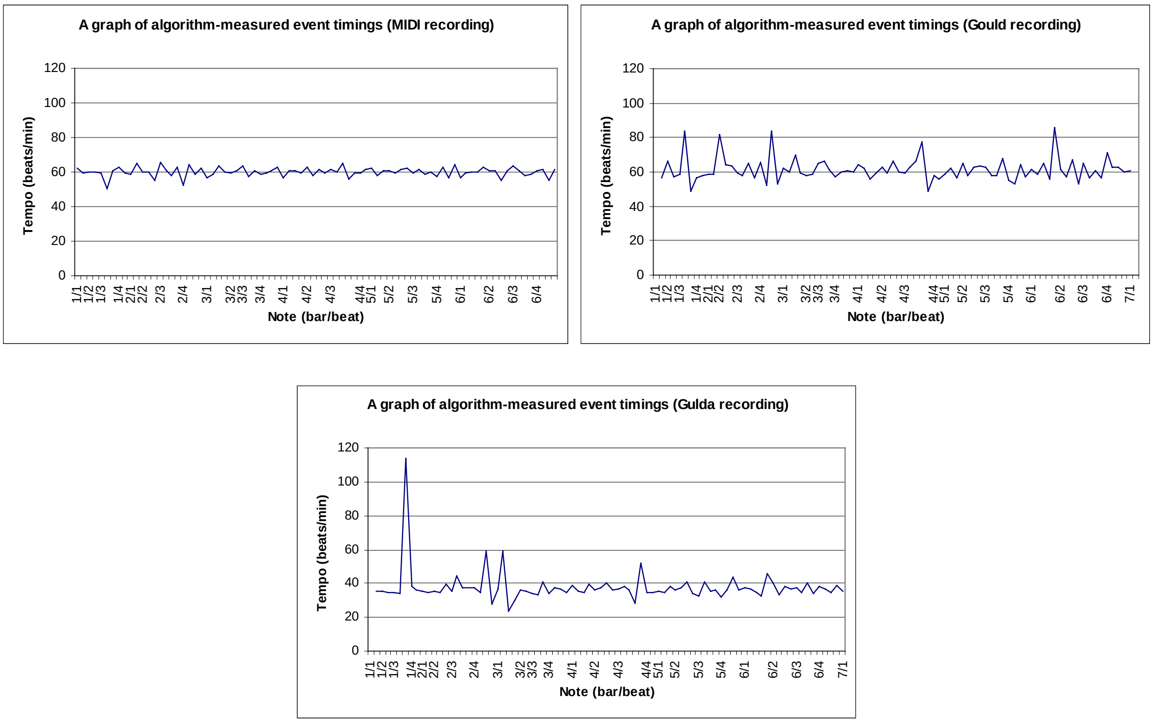


Figure 6. Tempo measurements in the opening bars of three performances of Bach's Fugue in C major, BWV 846b, with algorithm-measured note onset times illustrated by vertical lines.

The note duration, or inter-onset interval (IOI), of each note was computed. Graphs displaying this timing information, similar in format to those of Repp[5] and Gabrielsson[17], are illustrated in Figure 6. However, instead of using the absolute timing values, we present them in terms of an instantaneous measure of the number of beats per minute, referred to as instantaneous tempo, calculated from the individual IOIs. Table 1 shows the average tempo of each of the three extracts, together with the standard deviation of the tempo. The relative (or percentage) standard deviation of the instantaneous tempo is used as a measure of the overall amount of rubato. This is referred to as the 'rubato number'.

	MIDI	Gould	Gulda
Mean instantaneous tempo (beats/min)	60.01	61.56	37.82
Standard deviation	2.67	6.89	9.86
Rubato number (%)	**4.5**	**11.2**	**30.0**

Table 1. The mean instantaneous tempo and rubato number of note timings for the opening bars of three recordings of Bach's Fugue in C major BWV 846b

From these results, it can be seen that the performance by Gulda is both the slowest and exhibits the most rubato.

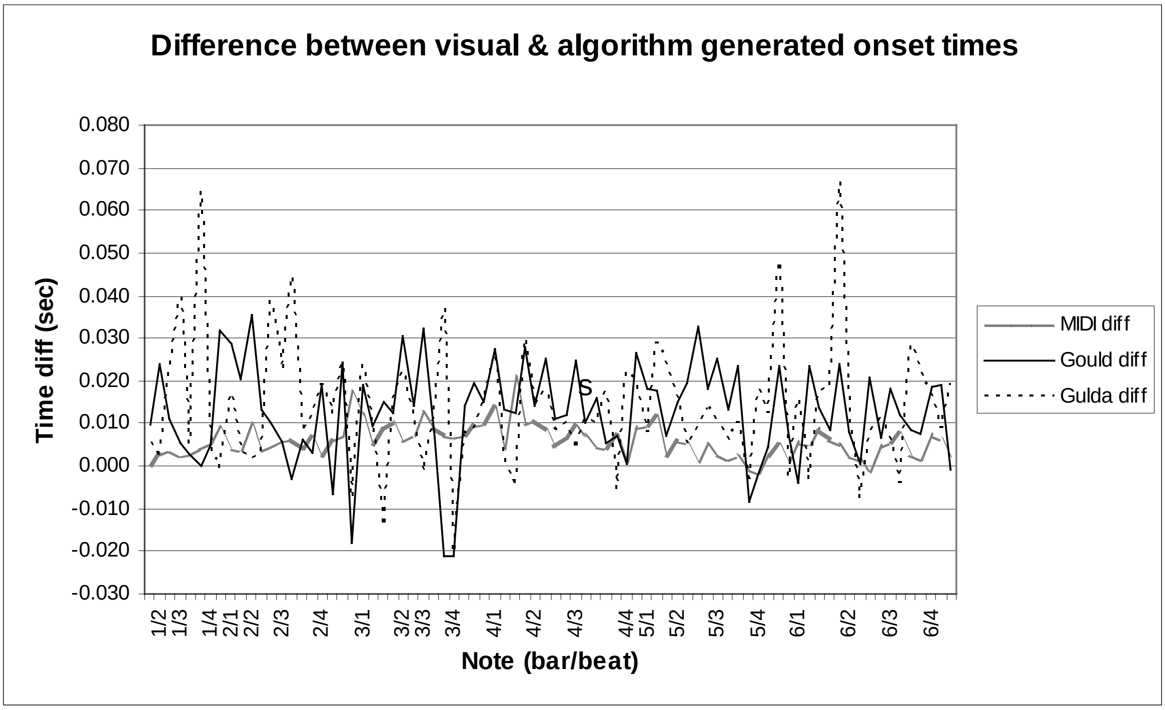


Figure 7. A comparison of manual and algorithm-measured note onset times.

For each of these performances, the note onset times were also measured manually, by studying the waveforms using visual and auditory clues. The accuracy of the note onset detection algorithm was measured by comparison of the manual and algorithm generated onset measurements. A graph of these results is given in Figure 7. These results are summarized in Table 2.

	MIDI	Gould	Gulda
Mean difference (ms)	5.8	12.6	13.7
Standard deviation (ms)	3.9	11.7	15.0

Table 2. The difference between manual and algorithm-generated note onset timings for the opening bars of three recordings of Bach's Fugue in C major BWV 846b

Algorithm measured event dynamics are illustrated in Figure 8. In each case, the results are scaled such that the mean dynamics (volume) is 1.0.

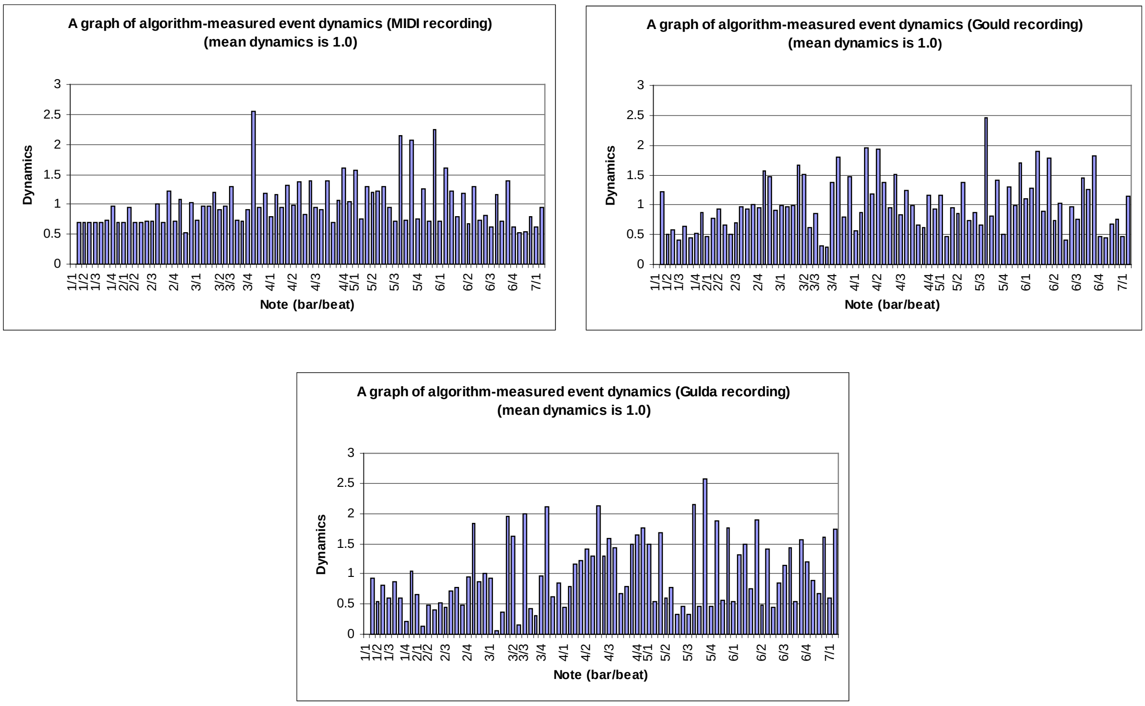


Figure 8. Algorithm-measured event dynamics of the opening bars of three performances of Bach's Fugue in C major, BWV 846b. The graphs show the variation from the average energy for all measured events.

The overall dynamic variation in the three recordings can be measured by looking at the standard deviation of the individual event dynamics. The relative (or percentage) standard deviation in dynamics of the MIDI, Gould and Gulda recordings respectively are 39.6, 45.1 and 57.1 percent. The performance by Gulda exhibits the most dynamic variation.

5 CONCLUSIONS AND FUTURE WORK

An algorithm has been presented to extract expressive timing and dynamics information from acoustic recordings of piano music, using the discretised continuous wavelet transform with a wavelet shape characteristic of that of the initial decay of individual piano note harmonics. Comparison of manual and algorithm measured note onset times gives encouraging results. Expressive dynamics are also extracted and compared. Future work involves the testing and optimisation of this wavelet on a wider range of acoustic recordings of piano music, and using the wavelet to measure other expressive performance parameters including articulation and pedalling. In this paper, a wavelet shape with a single set of optimization parameters, based on the mean values measured from a number of acoustically recorded individual piano notes, has been used. Extensions to this work may involve the optimization of these wavelet parameters for each particular note being measured.

6 REFERENCES

1. A. Binet, and J. Courtier, “Recherches graphiques sur la musique,” *L’Année Psychologique*, 2, pp. 201-222 (1895).
2. Seashore, C.E., *Psychology of music*, New York: McGraw-Hill (Reprinted 1967 by Dover Publications, New York). (1938).
3. W.L. Windsor, & E.F. Clarke, “Expressive timing and dynamics in real and artificial musical performance: using an algorithm as an analytical tool,” *Music Perception*, 15, pp. 127-152. (1997).
4. B.H. Repp, “Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists,” *J. Acous. Soc. Am.*, 88, pp. 622-650. (1990).
5. B.H. Repp, “A microcosm of musical expression: I. Quantitative analysis of pianists’ timing in the initial measures of Chopin’s Etude in E major,” *J. Acous. Soc. Am.*, 104, pp. 1085-1100. (1998).
6. B.H. Repp, “A microcosm of musical expression: II. Quantitative analysis of pianists’ dynamics in the initial measures of Chopin’s Etude in E major,” *J. Acous. Soc. Am.*, 105, pp. 1972-1988. (1999).
7. S. Dixon, “Automatic extraction of tempo and beat from expressive performances,” *Journal of New Music Research* 30, pp. 39-58. (200)1.
8. E.D. Scheirer, “Tempo and beat analysis of acoustic musical signals,” *J. Acoust. Soc. Am.*, 103, pp. 588-601. (1998).
9. S. Dixon, “Towards automatic analysis of expressive performance,” *5th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM5),* 8 - 13 September 2003, Hanover, Germany, pp 107-110. (2003).
10. O. Rioul, & M. Vetterli, “Wavelets and Signal Processing,” *Signal Processing Magazine*, IEEE, 8, pp. 14-38. (1991).
11. B.H. Repp, “Some empirical observations on sound level properties of recorded piano tones,” *J. Acoust. Soc. Am*, 93, pp. 1136-1144. (1993).
12. D.W. Martin, “Decay rates of piano tone,” *J. Acoust. Soc. Am.*, 19, 535-541. (1947).
13. B.E. Richardson, “Cambridge Companion to the Piano”, Ed. D. Rowland, Cambridge University Press, 96-113. (1998).
14. G. Weinreich, “The coupled motion of piano strings”, *Scientific American*, 240, 94-102. (1979).
15. G. Weinreich, “Coupled piano strings”, *J. Acoust. Soc. Am.*, 62, 1474-1484. (1977).
16. Wogram, K., "Acoustical Research on Pianos: Vibrational Characteristics of the Soundboard," *Das Musikinstrument*, 24, 694-702, 776-782, 872-880. (1980).
17. A. Gabrielsson, “The performance of music”, in The Psychology of Music (2nd edn) Ed. D. Deutsch, Academic Press, 501-602. (1998).