A Study of Automatic Transcription of Music Using a Standard PC

Victor Poon

2001

Abstract

This thesis describes using a Personal Computer to identify notes that are played by a musical instrument. Several groups have been doing this work with more sophisticated laboratories and equipment with only moderate success.

We have found the waves created by musical instruments vary, between instruments, a great deal in their stability and inherent vibration. It was more difficult to identify notes with very low frequencies than those with more central frequencies. We found it was very important to choose the correct starting point for the analysis with Fourier Transform otherwise we would not be analysing the stable stage of the wave.

We tried simple strategies to initially reduce the number of computer operations, and memory requirements, with marginal success. We followed with more complex subtraction strategies which were much more successful. The most useful technique involved creating a "calculated percentage multiple" which was almost 100% successful in identifying single notes. For multiple notes we were surprised to find that a group of five different instruments from MIDI were a better source of "known" notes to compare with the "unknown" notes than the MIDI equivalent of the real instrument playing the music.

These methods were developed using midi instruments but were verified using a real grand piano.

We suggest some further lines of enquiry that may make this technique more successful.

Acknowledgement

I would like to express my thanks to my supervisor, Dr. Peter Kay, for his help and advice during the completion of this thesis.

I am also very grateful for the help and the time given by Dr. Murray Sampson and Mrs. Lorraine Sampson who patiently proof read this thesis.

Finally I would like to thank my parents Yu Lin Poon and Tit Wing Poon for their financial and family support during my university study.

Contents

List of Figures

Chapter Three: Real Instruments

Appendix B

List of Tables

Chapter One: Background

Chapter Two: Using MIDI To Analyse Sound

Chapter Three: Real Instruments

Appendix B

Appendix C

Here we provide a CD which contains the programs we wrote to enable us to perform the experiments and analyses.

Introduction

This thesis describes the work we have done to study the automatic transcription of music to printed notation using a standard PC. After the subject of this thesis was accepted we began a literature review. This review showed we would need to reduce the scope of the work to enable us to make a valid contribution to the field. We have therefore limited the study to identifying notes by comparing them to MIDI and also notes from the same instrument. Later in our work we realised that the selection of the starting point for the comparison was very important. However the time available did not allow us to more than touch on the issue. This complexity has also created difficulties for other work (Hamer 2001 [16]) done by larger groups using more sophisticated equipment.

We quote from another worker "Our scope and treatise is limited by several factors, but especially by the limited amount of resources compared to the wide range of topics that are related with music transcription. Moreover, engaging in a research that is quite new to our laboratory, analysis of musical signals, called for paying the required attention to just finding the right points of emphasis and avoiding wrong assumptions in an early phase." (Klapuri 2002 [21]) The quoted author is one member of a group six. We quote further discussion on commercial products "Polyphonic transcribers – sad to say – work very poorly. Monophonic systems are more robust, of course."

Our research has shown similar difficulties to those of other workers and taking into account the limitations, our accuracy has been acceptable.

Chapter 1 gives the background to the basic techniques used in this thesis. In sections one to four, we will explain the music and file systems we used. In sections five and six, we will look at the advantages of using MATLAB and the use of Fourier Transform.

Chapter 2 explains and reports on the analysis of MIDI generated files. It covers the analysis on single and a small number of notes played together.

Chapter 3 will perform the similar analysis on a real piano.

We then report our conclusions.

Literature Review

When one examines the literature, it becomes clear that the mathematical study of music is a difficult field studied over many years and from many different aspects. Such studies are the basis for the ability to transcribe music using a PC. In this review we give a general overview with reference examples rather than extensive lists. Our other approach is to cite reviews, textbooks and homepages to enable the reader to home in on any particular interest they may have.

When the unaware see sheet music they would be inclined to believe that what they see tells them what they will hear, if it is played by a competent musician. Unfortunately the simple indication of the note to be played does not reveal the way the music is created by instrument or voice. The sound produced varies a great deal depending on the instrument. The pitch is the same but the timbre is different.

Pitched instruments may make their sound through mechanisms that depend on creating a vibration in a physical object and modifying it before it transmits through air to the audience. A "string" maybe struck as in a piano or "rubbed' as in a violin. The length of the string, either in the instrument, or modified by "fingering", will set the frequency of the vibration, hence the sound. There will usually be more than one frequency created as part of one "note". Further modification of sound is via the construction material of the instrument, through hollow spaces such as in the body of a violin, or sounding boards in a piano.

Wind instruments make sound through the use of air forced through by either special oral techniques or mechanisms such as reeds to cause vibration. Their sound is affected by the length of the tube the air passes through. Most Brass instruments shorten the length of the tube by using valves to close off part of the tube. Other wind instruments such as the flute, depend on shortening the amount of tube that can allow air vibration, by covering the holes in the tube. Shortening the tube length alters the frequency of the note. see Suggs (1966) [35]. The physics of musical instruments is very complicated and is explained in Fletcher and Rossing (1998) [10]. This reference may be the prior edition of Rossing, Wheeler and Moore (2002) [34] but is sufficiently different to be worth referring to separately.

These differences, and others, are the source of the complications found in the automatic transcription of music. Notes may "overlap" or be played together.

The science of sound is explained in Rossing, Wheeler, Moore (2002) [34].

A recent review of mathematics and music is Benson (2002) [3]. This is an expanded and augmented version of notes given for an undergraduate course at the University of Georgia. Water waves are waves where the local movement is at right angles to the direction of propagation (transverse waves) whereas sound waves are longitudinal waves with the local movements in the same direction as the propagation. Sound waves have four main attributes:

1. Amplitude is the size of the vibration and is perceived as loudness. The amplitude of typical everyday sound is only a small fraction of a millimetre.
2. Pitch corresponds to the frequency of the vibration.
3. Timbre corresponds to the shape of the frequency spectrum of the sound, and
4. Duration which is the length of time for which the note lasts.

However most vibrations do not consist of a single frequency and naming the "defining" frequency can be difficult. Music should be defined more in terms of the perception of sound rather than in terms of the sound itself. The perceived pitch of a sound may represent a frequency not actually present in the waveform -- a "missing fundamental". This phenomenon is a part of "Psychoacoustics".

Benson describes the functioning of the human ear and its limitations. It "hears within a range of about 20Hz (Hertz) to 20,000 Hz although it may "feel" the sound below 20Hz.

The relevance of sine waves in relation to sound is explained as lying in the differential equation for simple harmonic motion which can be taken as a close approximation to the equation of motion of a particular point on the basilar membrane of the human ear or anywhere else on the chain from the outside air to the cochlea. The limitations of the approximation are discussed. Harmonic motion is explained in mathematical terms. This leads to damped harmonic motion and resonance.

Refer to Benson (2002) [3] where he then discusses Fourier theory and a useful mathematical system, which is often used for the analysis of sound, being Fourier Transform.

The article usually regarded as the original, announcing the fast Fourier Transform as a practical algorithm, is Cooley and Tukey (1965) [6]. An early algorithm is given by Gold and Rabiner (1969) [12]. This algorithm has been used much later in New Zealand to develop an initial system for the identification of melodies. McNab, Smith, Bainbridge and Witten (1997) [25]. An earlier description of signal processing, which involves measurement of spectra using Fourier transform is given in the text, Rabiner and Gold (1975) [31].

The creation of music by instrument is one way of starting the chain. The other is the voice. Some studies are involved in relating sound produced to the characteristics of the larynx. An example is Bagshaw, Hiller and Jack (1993) [1]. The next link is the transmission of sound through air and its subsequent interaction with objects within the space or enclosing the space. These matters of acoustics are of importance in the transcription of music because they indicate the possibility of the sound being changed before it is identified. (Kinsler, L.E, Frey, Coppens, A.B. , and Sanders 1982 [19]). The shape of the room in which it is played, and the nature of the floor and its foundations, are two examples of the factors which affect the acoustics.

The human perception of music is another aspect which is also very complicated (Roederer 1995 [33]). The functioning of the human ear is an integral part of this perception. (Hudspeth 1985 [17], 1989 [18]) is another reference for this aspect. Further references are continually updated on the homepage (CSTR Publications 2001 [7]). Neuropsychology is the discipline that studies the neural system processes and functions linking the input received from the environment and body with the full behavioural and mental output. A text covering this is McAdams and Bigand (1993) [24]. A homepage that covers this field is Neuropsychology Central (2001) [29].

The transcription of music involves artificial perception and music recognition rather than human perception. Tanguiane (1993) [36] gives us a good overview of this topic. Work dates back to the early 1970's with the first experiments on automatic notation of monophonic music and a program for transcribing polyphonic music performed on a computer-wired keyboard. The limitations experienced at that time included only admitting pitched sounds, avoiding some pitch combinations, and requiring simple rhythmic structure and a constant tempo. Artificial intelligence methods began to be used in the early to the mid 1980's.

Artificial perception is the usual approach to the pattern recognition which is necessary to develop a computer system for the automatic notation of the performed music. It attempts to follow nature. A difficulty is that if a system of artificial perception is developed for one type of music, it may not be as successful with different styles or musical cultures. Another problem is that compared to human perception, a much longer passage of music is required for artificial perception to establish a reliable result. A singing voice, being less stable than an instrumental sound, is more difficult to analyse. By 1985 the studies branch into either chord/note recognition or rhythm/tempo recognition.

An early method for rhythm recognition was to develop an hypothesis concerning the rhythmic structure from the very first events which is then continuously confronted with current data and being modified if necessary.

We paraphrase a passage that states that a particular difficulty is that there are no explicit definitions of notes, chords, rhythm, and tempo, and so their recognition is complicated. Goto and Muraoka (1997) [14] discuss the issue of the evaluation of beat tracking systems and the need to rely on human intervention as an evaluation tool since the beat is a perceptual concept a person feels in music which is difficult to define in an objective way. In designing their measure for evaluation they considered subjective hand-labeled beat positions to be the correct beat times. The positions can be finely adjusted by playing back the audio with click tones at beat times and the user also defines a hierarchical rhythmic structure. They later (Goto and Muraoka, 1998 [15]) report on a real-time beat tracking system that recognises a rhythmic structure in realworld audio signals sampled from popular-music compact discs. A homepage as an entry into the literature on rhythm/tempo recognition is Goto (2002) [13].

A recent method (Klapuri, Virtanen and Holm 2000 [22]) for the estimation of the multiple pitches of concurrent musical sounds comprised of sung vowels and the whole pitch range of 26 musical instruments had error rates for mixtures ranging from one to six simultaneous sounds were 2.1%, 2.4%, 3.8%, 8.1%, 12% and 18% respectively. In musical interval and chord identification tasks, the algorithm outperformed the average of ten trained musicians. This gives an indication of the current accuracy being achieved.

It is interesting to realise that music recognition and voice recognition use the same methods and the studies overlap. A problem in common is the separation of several sounds being heard by the recognition system some of which are "interference" outside the voice or music being studied. A recent review is De Cheveigne (1993) [8]. Results from an implementation of this approach illustrated its ability to analyse complex, ambient sound scenes that would confound previous systems.

Homepages for groups working in this field include MIT Media Lab (2002) [27] and Klapuri (2002) [21]. Because these are updated continually, they are ongoing sources of recent work also giving references to other workers.

A system of artificial perception, which has some similarity to artificial intelligence is the use of neural network. Examples of this concept are given in Roberto (1993) [32]. Other work with neural networks can be accessed at CSTR publications (2001) [7].

The review to this stage showed that it was going to be very difficult for us to create a research project within the limit of time and facilities. Klapuri (2002) [21] mentions a similar problem. It is interesting to note that Klapuri made this statement after completing the thesis (Klapuri 1998 [20]) within Tampere University of Technology, Finland, which has an Audio Research group. Klapuri (1998) [20] gives a review of systems of artificial perception. The phrase onset time is defined as the instant of time when the sound starts playing.

An important distinction made by Tanguiane (1993) [36] is quoted.

"The difference between artificial perception and artificial intelligence in pattern recognition is understood as follows. Artificial perception is used for discovering structure in visual and audio images by self-organisation of data and segregation of patterns. Artificial intelligence is used for pattern identification by their matching to known concepts. Usually, the identification of already segregated patterns is much simpler than their recognition in data flows: thus artificial perception and artificial intelligence are complementary."

This distinction by Tanguiane (1993) [36] is not necessarily clear in the use of the phrase, artificial intelligence, by other workers, but it leads to another approach which is used in our studies.

In August 1983, music manufacturers agreed on a protocol that is called the "MIDI 1.0 Specification". General MIDI (Musical Instrument Digital Interface) is a standard that defines specific locations for different instrument sounds in the present memory of synthesisers (MIDI Manufacturers Association Incorporated 2001 [26]).

Therefore MIDI could be used as the known concept, referred to by Tanguiane (1993) [36] to develop a system of musical note identification. Benson (2002) [3] gives a recent review of concepts in digital music.

Work which has some similarities to ours has used matching methods to compare a hummed tune to a recorded database of songs. We refer to Ghias and Logan (1995) [11] and Blackburn and DeRoure (1998) [4]. They have used a pitch tracking method represented as a sequence of relative pitch changes (i.e a melodic pitch contour). We note that when a person hums a song it is monophonic.

The New Zealand Digital Library MELody inDEX (McNab, Smith. Bainbridge and Witten (1997) [25]) is a system that accepts acoustic input from the user created by a few notes sung into the microphone. It transcribes the melodies automatically from the microphone input then searches a database for tunes that contain the same or similar sung patterns. The tunes retrieved are ranked according to the closeness of the match. Different search criteria were used such as melodic contour, musical intervals and rhythm, and tests were performed using both exact and approximate string matching. They also performed tests on how people remember tunes. They concluded from these experiments that people needed a choice of several matching procedures and should be able to explore the results interactively in their search for a particular melody. This recent effort shows the recurring need for human interaction in these processes. The voice range was limited to from F2 to G5. This reference is dealing with monophonic sound and uses the algorithm on Gold and Rabiner (1969) [12]. The system depends on the user separating each note by singing da or ta to create a note boundary. The solutions to the other problems and the techniques used are described. On going access to the work of this group can be made via the homepage of Bainbridge (2002) [2].