Category: Multimedia Technology

Audio Analysis Applications for Music

Simon Dixon
Austrian Research Institute for Artificial Intelligence, Austria

INTRODUCTION

The last decade has seen a revolution in the use of digital audio: The CD, which one decade earlier had taken over the home audio market, is starting to be replaced by electronic media which are distributed over the Internet and stored on computers or portable devices in compressed formats. The need has arisen for software to manage and manipulate the gigabytes of data in these music collections, and with the continual increase in computer speed, memory and disk storage capacity, the development of many previously infeasible applications has become possible.

This article provides a brief review of automatic analysis of digital audio recordings with musical content, a rapidly expanding research area which finds numerous applications. One application area is the field of music information retrieval, where content-based indexing, classification and retrieval of audio data are needed in order to manage multimedia databases and libraries, as well as being useful in music retailing and commercial information services. Another application area is music software for the home and studio, where automatic beat tracking and transcription of music are much desired goals. In systematic musicology, audio analysis algorithms are being used in the study of expressive interpretation of music. Other emerging applications which make use of audio analysis are music recommender systems, playlist generators, visualisation systems, and software for automatic synchronisation of audio with other media and/or devices.

We illustrate recent developments with three case studies of systems which analyse specific aspects of music (Dixon, 2004). The first system is BeatRoot (Dixon, 2001a, 2001c), a beat tracking system that finds the temporal location of musical beats in an audio recording, analogous to the way that people tap their feet in time to music. The second system is JTranscriber, an interactive automatic transcription system based on (Dixon, 2000a, 2000b), which recognizes musical notes and converts them into MIDI format, displaying the audio data as a spectrogram with the MIDI data overlaid in piano roll notation, and allowing interactive monitoring and correction of the extracted MIDI data. The third system is the Performance Worm (Dixon, Goebl, & Widmer, 2002), a real-time system for visualisation of musical expression, which presents in real time a two dimensional animation of variations in tempo and loudness (Langner & Goebl, 2002, 2003).

Space does not permit the description of the many other music content analysis applications, such as: audio fingerprinting, where recordings can be uniquely identified with a high degree of accuracy, even with poor sound quality and in noisy environments (Wang, 2003); music summarisation, where important parts of songs such as choruses are identified automatically; instrument identification, using machine learning techniques to classify sounds by their source instruments; and melody and bass line extraction, essential components of query-by-example systems, where music databases can be searched by singing or whistling a small part of the desired piece. At the end of the article, we discuss emerging and future trends and research opportunities in audio content analysis.

BACKGROUND

Early research in musical audio analysis is reviewed by Roads (1996). The problems that received the most attention were pitch detection, spectral analysis and rhythm recognition, areas which correspond respectively to the three most important aspects of music: melody, harmony and rhythm.

Pitch detection is the estimation of the fundamental frequency of a signal, usually assuming it to be monophonic. Methods include: time domain algorithms such as counting of zero-crossings and autocorrelation; frequency domain methods such as Fourier analysis and the phase vocoder; and auditory models which combine time and frequency domain information based on an understanding of human auditory processing. Recent work extends these methods to find the predominant pitch (e.g., the melody note) in a polyphonic mixture (Gómez, Klapuri, & Meudic, 2003; Goto & Hayamizu, 1999).

Spectral analysis is a well-understood research area with many algorithms available for analysing various classes of signals, such as the short time Fourier transform, wavelets and other more signal-specific time-frequency distributions. Building upon these methods, the specific application of automatic music transcription has a long research history (Chafe, Jaffe, Kashima, Mont-Reynaud, & Smith, 1985; Dixon, 2000a, 2000b; Kashino, Nakadai, Kinoshita, & Tanaka, 1995; Klapuri, 1998, 2003; Klapuri, Virtanen, & Holm, 2000; Marolt, 1997, 1998, 2001; Martin, 1996; Mont-Reynaud, 1985; Moorer, 1975; Piszczalski & Galler, 1977; Sterian, 1999; Watson, 1985). Certain features are

common to many of these systems: producing a time-frequency representation of the signal, finding peaks in the frequency dimension, tracking these peaks over the time dimension to produce a set of partials, and combining the partials to produce a set of notes. The differences between systems are usually related to the assumptions made about the input signal (e.g., the number of simultaneous notes, types of instruments, fastest notes, or musical style), and the means of decision making (e.g., using heuristics, neural nets or probabilistic reasoning).

The problem of extracting rhythmic content from a musical performance, and in particular finding the rate and temporal location of musical beats, has also attracted considerable interest in recent times (Allen & Dannenberg, 1990; Cemgil, Kappen, Desain, & Honing, 2000; Desain, 1993; Desain & Honing, 1989; Dixon, 2001a; Eck, 2000; Goto & Muraoka, 1995, 1999; Large & Kolen, 1994; Longuet-Higgins, 1987; Rosenthal, 1992; Scheirer, 1998; Schloss, 1985). Previous work had concentrated on rhythmic parsing of musical scores, lacking the tempo and timing variations that are characteristic of performed music, but recent tempo and beat tracking systems work quite successfully on a wide range of performed music.

Music performance research is only starting to take advantage of the possibility of audio analysis software, following work such as Scheirer (1995) and Dixon (2000a). Previously, general purpose signal visualisation tools combined with human judgement had been used to extract performance parameters from audio data. The main problem in music signal analysis is the development of algorithms to extract sufficiently high level content, since it requires the type of musical knowledge possessed by a musically literate human listener. Such "musical intelligence" is difficult to encapsulate in rules or algorithms that can be incorporated into computer programs. In the following sections, three systems are presented which take the approach of encoding as much as possible of this intelligence in the software and then presenting the results in an intuitive format which can be edited via a graphical user interface, so that the systems can be used in practical settings even when not 100% correct. This approach has proved to be very successful in performance research (Dixon et al., 2002; Goebl & Dixon, 2001; Widmer, 2002; Widmer, Dixon, Goebl, Pampalk, & Tobudic, 2003).

BEATROOT

Compared with complex cognitive tasks such as playing chess, beat tracking (identifying the basic rhythmic pulse of a piece of music) does not appear to be particularly difficult, as it is performed by people with little or no musical training, who tap their feet, clap their hands or dance in time with music. However, while chess programs compete with world champions, no computer program has been developed which approaches the beat tracking ability of an average musician, although recent systems are approaching this target. In this section, we describe BeatRoot, a system which estimates the rate and times of musical beats in expressively performed music (for a full description, see Dixon, 2001a, 2001c).

BeatRoot models the perception of beat by two interacting processes (see Figure 1): The first finds the rate of

Figure 1. System architecture of BeatRoot