Automatic Music Timbre Indexing

Xin Zhang
University of North Carolina at Charlotte, USA

Zbigniew W. Ras
University of North Carolina, Charlotte, USA

INTRODUCTION

Music information indexing based on timbre helps users to get relevant musical data in large digital music databases. Timbre is a quality of sound that distinguishes one music instrument from another among a wide variety of instrument families and individual categories. The real use of timbre-based grouping of music is very nicely discussed in (Bregman, 1990).

Typically, an uncompressed digital music recording, in form of a binary file, contains a header and a body. A header stores file information such as length, number of channels, rate of sample frequency, etc. Unless being manually labeled, a digital audio recording has no description on timbre, pitch or other perceptual properties. Also, it is a highly nontrivial task to label those perceptual properties for every piece of music object based on its data content. Lots of researchers have explored numerous computational methods to identify the timbre property of a sound. However, the body of a digital audio recording contains an enormous amount of integers in a time-order sequence. For example, at a sample frequency rate of 44,100Hz, a digital recording has 44,100 integers per second, which means, in a one-minute long digital recording, the total number of the integers in the time-order sequence will be 2,646,000, which makes it a very big data item. Being not in form of a record, this type of data is not suitable for most traditional data mining algorithms.

Recently, numerous features have been explored to represent the properties of a digital musical object based on acoustical expertise. However, timbre description is basically subjective and vague, and only some subjective features have well defined objective counterparts, like brightness, calculated as gravity center of the spectrum. Explicit formulation of rules of objective specification of timbre in terms of digital descriptors will formally express subjective and informal sound characteristics. It is especially important in the light of human perception of sound timbre. Time-variant information is necessary for correct classification of musical instrument sounds because quasi-steady state, where the sound vibration is stable, is not sufficient for human experts. Therefore, evolution of sound features in time should be reflected in sound description as well. The discovered temporal patterns may better express sound features than static features, especially that classic features can be very similar for sounds representing the same family or pitch, whereas changeability of features with pitch for the same instrument makes sounds of one instrument dissimilar. Therefore, classical sound features can make correct identification of musical instruments independently on the pitch very difficult and erroneous.

BACKGROUND

Automatic content extraction is clearly needed and it relates to the ability of identifying the segments of audio in which particular predominant instruments were playing. Instruments having rich timbre are known to produce overtones, which result in a sound with a group of frequencies in clear mathematical relationships (so-called harmonics). Most western instruments produce harmonic sounds. Generally, identification of musical information can be performed for audio samples taken from real recordings, representing waveform, and for MIDI (Musical Instrument Digital Interface) data. MIDI files give access to highly structured data. So, research on MIDI data may basically concentrate on higher level of musical structure, like key or metrical information. Identifying the predominant instruments, which are playing in the multimedia segments, is even more difficult. Defined by ANSI as the attribute of auditory sensation, timbre is rather subjective: a quality of sound, by which a listener can judge that two sounds, similarly presented and having the same loudness and pitch, are different. Such definition is subjective and not of much use for automatic sound timbre classification. Therefore, musical sounds must be very carefully parameterized to allow automatic timbre recognition. There are a number of different approaches to sound timbre (Balzano, 1986; Cadoz, 1985). Dimensional approach to timbre description was proposed by (Bregman, 1990). Sets of acoustical features have been successfully developed for timbre estimation in monophonic sounds where mono instruments were playing. However, none of those features can be successfully applied to polyphonic sounds, where two or more instruments were playing at the same time, since those features represent the overlapping sound harmonics as a whole instead of individual sound sources.

This has brought the research interest into Blind Source Separation (BBS) and independent component analysis (ICA) for musical data. BBS is to estimate original sound sources based on signal observations without any knowledge on the mixing and filter procedure. ICA is to separate sounds by linear models of matrix factorization based on the assumption that each sound source is statistically independent. Based on the fact that harmonic components have significant energy, harmonics tracking together with Q-Constant Transform and Short Time Fourier Transform have been applied to sound separation (Dziubinski, Dalka and Kostek 2005; Herrera, Peeters and Dubnov 2003; Zhang and Ras 2006B). The main steps in those researches include processing polyphonic sounds into monophonic sounds, extracting features from the resultant monophonic sounds, and then performing classification.

MAIN FOCUS

Current research in timbre recognition for polyphonic sounds can be summarized into three steps: sound separation, feature extraction and classification. Sound separation has been used to process polyphonic sounds into monophonic sounds by isolating sound sources; features have been used to represent the sound behaviors in different domains; then, classification shall be performed based on the feature values by various classifiers.

Sound Separation

In a polyphonic sound with multiple pitches, multiple sets of harmonics from different instrument sources are overlapping with each other. For example, in a sound mix where a sound in 3A of clarinet and a sound in 4C of violin were played at the same time, there are two sets of harmonics: one set is distributed near several integer multiples of 440Hz; the other spreads around integer multiples of 523.25Hz. Thus, the j^{th} harmonic peak of the k^{th} instrument can be estimated by searching a local peak in the vicinity of an integer multiple of the fundamental frequency. Consequently, k predominant instruments will result in k sets of harmonic peaks. Then, we can merge the resultant sets of harmonic peaks together to form a sequence of peaks H_p^j in an ascending order by the frequency, where three possible situations should be taken into consideration for each pair of neighbor peaks: the two immediate peak neighbors are from the same sound source; the two immediate peak neighbors are from two different sound sources; part of one of the peak and the other peak are from the same sound source. The third case is due to two overlapping peaks, where the frequency is the multiplication of the fundamental frequencies of two different sound sources. In this scenario, the system first partitions the energy between the two sound sources according to the ratio of the previous harmonic peaks of those two sound sources. Therefore, only the heterogeneous peaks should be partitioned. A clustering algorithm has been used for separation of energy between two immediate heterogeneous neighbor peaks. Considering the wide range of the magnitude of harmonic peaks, we may apply a coefficient to linearly scale each pair of immediate neighbor harmonic peaks to a virtual position along the frequency axis by a ratio of the magnitude values of the two harmonic peaks. Then the magnitude of each point between the two peaks is proportionally computed in each peak. For fast computation, a threshold for the magnitude of each FFT point has been applied, where only points with significant energy had been computed by the above formulas. We assume that a musical instrument is not predominant only when its total harmonic energy is significantly smaller than the average of the total harmonic energy of all sound sources. After clustering the energy, each FFT point in the analysis window has been assigned k coefficients, for each predominant instrument accordingly.