Music Information Retrieval

Alicja A. Wieczorkowska
Polish-Japanese Institute of Information Technology, Poland

INTRODUCTION

Music information retrieval is a multi-disciplinary research on retrieving information from music. This research involves scientists from traditional, music, and digital libraries; information science; computer science; law; business; engineering; musicology; cognitive psychology; and education (Downie, 2001).

BACKGROUND

A huge amount of audio resources, including music data, is becoming available in various forms, both analog and digital. Notes, CDs, and digital resources of the World Wide Web are growing constantly in amount, but the value of music information depends on how easy it can be found, retrieved, accessed, filtered, and managed (Fingerhut, 1997; International Organization for Standardization, 2003).

Music information retrieval consists of quick and efficient searching for various types of audio data of interest to the user, filtering them in order to receive only the data items that satisfy the user's preferences (International Organization for Standardization, 2003). This broad domain of research includes:

- Audio retrieval by content;
- Auditory scene analysis and recognition (Rosenthal & Okuno, 1998);
- Music transcription;
- Denoising of old analog recordings; and
- Other topics discussed further in this article (Wieczorkowska & Ras, 2003).

The topics are interrelated, since the same or similar techniques can be applied for various purposes. For instance, source separation, usually applied in auditory scene analysis, is used also for music transcription and even restoring (denoising) of all recordings. The research on music information retrieval has many applications. The most important include automatic production of music score on the basis of the presented input, retrieval of music pieces from huge audio databases, and restoration of old recordings. Generally, the research within the music information retrieval domain is focused on harmonic structure analysis, note extraction, melody and rhythm tracking, timbre and instrument recognition, classification of type of the signal (speech, music, pitched vs. non-pitched), and so forth.

The research basically uses digital audio recordings, where sound waveform is digitally stored as a sequence of discrete samples representing the sound intensity at a given time instant, and MIDI files, storing information on parameters of electronically synthesized sounds (voice, note on, note off, pitch bend, etc.). Sound analysis and data mining tools are used to extract information from music files in order to provide the data that meet users' needs (Wieczorkowska & Ras, 2001).

MAIN THRUST

Music information retrieval domain covers a broad range of topics of interest, and various types of music data are investigated in this research. Basic techniques of digital sound analysis for these purposes come from speech processing focused on automatic speech recognition and speaker identification (Foote, 1999). Sound descriptors calculated this way can be added to the audio files in order to facilitate content-based searching of music databases.

The issue of representation of music and multimedia information in a form that allows interpretation of the information's meaning is addressed by MPEG-7 standard, named Multimedia Content Description Interface. MPEG-7 provides a rich set of standardized tools to describe multimedia content through metadata (i.e., data about data), and music information description also has been taken into account in this standard (International Organization for Standardization, 2003).

The following topics are investigated within music information retrieval domain:

- Auditory scene analysis, which focuses on various aspects of music like timbre description, sound harmonicity, spatial origin, source separation, and so forth (Bregman, 1990). Timbre is defined subjectively as this feature of sound that distinguishes two sounds of the same pitch, loudness, and duration. Therefore, subjective listening tests are often performed in this research, but also signal-processing techniques are broadly applied here. One of the main topics of computational auditory scene analysis is automatic separation of individual sound sources from a mixture. It is difficult with mixtures of harmonic instrument sounds, where spectra overlap. However, assuming time-frequency smoothness of the signal, sound separation can be performed, and when sound changes in time are observed, onset, offset, amplitude, and frequency modulation have similar shapes for all frequencies in the spectrum; thus, a demixing matrix can be estimated for them (Virtanen, 2003; Viste & Evangelista, 2003). Audio source separation techniques also can be used to source localization for auditory scene analysis. These techniques, like independent component analysis, originate from speech recognition in cocktail party environment, where many sound sources are present. Independent component analysis is used for finding underlying components from multidimensional statistical data, and it looks for components that are statistically independent (Vincent et al., 2003).

- Computational auditory scene recognition aims at classifying auditory scenes into predefined classes, using audio information only. Examples of auditory scenes are various outside and inside environments, like streets, restaurants, offices, homes, cars, and so forth. Statistical and nearest neighbor algorithms can be applied for this purpose. In the nearest neighbor algorithm, the class (type of auditory scene, in this case) is assigned on the basis of the distance of the investigated sample to the nearest sample, for which the class membership is known. Various acoustic features, based on Fourier spectral analysis (i.e., mathematic transform, decomposing the signal into frequency components), can be applied to parameterize the auditory scene for classification purposes. Effectiveness of this research approaches 70% correctness for about 20 auditory scenes (Peltonen et al., 2002).
- Query-by-humming systems, which search melodic databases using sung queries (Adams et al., 2003). This topic represents audio retrieval by contents. Melody usually is quantized coarsely with respect to pitch and duration, assuming moderate singing abilities of users. Music retrieval system takes such an aural query (i.e., a motif or a theme) as input, and searches the database for the piece from which this query comes. Markov models, based on Markov chains, can be used for modeling musical performances. Markov chain is a stochastic process for which the parameter is discrete time values. In Markov sequence of events, the probability of future states depends on the present state; in this case, states represent pitch (or set of pitches) and duration (Birmingham et al., 2001). Query-by-humming is one of more popular topics within music information retrieval domain.
- Audio retrieval-by-example for orchestral music aims at searching for acoustic similarity in an audio collection, based on analysis of the audio signal. Given an example audio document, other documents in a collection can be ranked by similarity on the basis of long-term structure; specifically, the variation of soft and louder passages, determined from envelope of audio energy vs. time in one or more frequency bands (Foote, 2000). This research is a branch of audio retrieval by content. Audio query-by-example search also can be performed within a single document when searching for sounds similar to the selected sound event. Such a system for content-based audio retrieval can be based on a self-organizing feature map (i.e., a special kind of neural network designed by analogy with a simplified model of the neural connections in the brain and trained to find relationships in the data). Perceptual similarity can be assessed on the basis of spectral evolution in order to find sounds of similar timbre (Spevak & Polfreman, 2001). Neural networks also are used in other forms in audio information retrieval systems. For instance, time-delayed neural networks (i.e., neural nets with time delay inputs) are applied, since they perform well in speech recognition applications (Meier et al., 2000). One of applications of audio retrieval-by-example is searching for the piece in a huge database of music pieces with the use of so-called audio fingerprinting—technology that allows piece identification. Given a short passage transmitted, for instance, via car phone, the piece is extracted, and, most important, information also is extracted on the performer and title linked to this piece in the database. In this way, the user may identify the piece of music with very high accuracy (95%) only on the basis of a small recorded (possibly noisy) passage.
- Transcription of music, defined as writing down the musical notation for the sounds that constitute the investigated piece of music. Onset detection based on incoming energy in frequency bands and multi-pitch estimation based on spectral analysis may be used as the main elements of an automatic music transcription system. The errors in such a system may contain additional inserted notes, omissions, or erroneous transcriptions (Klapuri et al., 2001). Pitch tracking (i.e., estimation of pitch of note events in a melody or a piece of music) is often performed in many music information retrieval systems. For polyphonic music,

Related Content