Automatic Musical Instrument Sound Classification

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

INTRODUCTION

The aim of musical instrument sound classification is to process information from audio files by a classificatory system and accurately identify musical instruments playing the processed sounds. This operation and its results are called automatic classification of musical instrument sounds.

BACKGROUND

Musical instruments are grouped into the following categories (Hornbostel & Sachs, 1914):

- **Idiophones:** Made of solid, non-stretchable, sonorous material.
- **Membranophones:** Skin drums; membranophones, and idiophones are called percussion.
- **Chordophones:** Stringed instruments.
- **Aerophones:** Wind instruments: woodwinds (single-reed, double-reed, flutes) and brass (lip-vibrated)

Idiophones are classified according to the material, number of idiophones and resonators in a single instrument, and whether pitch or tuning is important. Subcategories include idiophones struck together by concussion (e.g., castanets), struck (gong), rubbed (musical glasses), scraped (washboards), stamped (hard floors stamped with tap shoes), shaken (rattles), and plucked (jew's harp).

Membranophones are classified according to their shape, material, number of heads, if they have snares, etc., whether and how the drum is tuned, and how the skin is fixed and played. Subcategories include drums (cylindrical, conical, barrel, hourglass, goblet, footed, long, kettle, frame, friction drum, and mirliton/kazoo).

Chordophones are classified with respect to the relationship of the strings to the body of the instrument, if they have frets (low bridges on the neck or body, where strings are stopped) or movable bridges, number of strings, and how they are played and tuned. Subcategories include zither, lute plucked, and bowed (e.g., guitars, violin), harp, lyre, and bow.

Aerophones are classified according to how the air is set in motion, mainly depending on the mouthpiece: blow hole, whistle, reed, and lip-vibrated. Subcategories include flutes (end-blown, side-blown, nose, globular, multiple), panpipes, whistle mouthpiece (recorder), single- and double-reed (clarinet, oboe), air chamber (pipe organs), lip-vibrated (trumpet or horn), and free aerophone (bullroarers) (SIL, 1999).

The description of properties of musical instrument sounds is usually given in vague subjective terms, like sharp, nasal, bright, and so forth, and only some of them (i.e., brightness) have numerical counterparts. Therefore, one of the main problems in this research is to prepare the appropriate numerical sound description for instrument recognition purposes.

Automatic classification of musical instrument sounds aims at classifying audio data accurately into appropriate groups representing instruments. This classification can be performed at instrument level, instrument family level (e.g., brass), or articulation (i.e., how sound is struck, sustained, and released, e.g. vibrato - varying the pitch of a note up and down) (Smith, 2000). As a preprocessing, the audio data are usually parameterized (i.e., numerical or other parameters or attributes are assigned, and then data mining techniques are applied to the parameterized data). Accuracy of classification varies, depending on the audio data used in the experiments, number of instruments, parameterization, classification, and validation procedure applied. Automatic classification compares favorably with human performance. Listeners identify musical instruments with accuracy far from perfect, with results depending on the sounds chosen and experience of listeners. Classification systems allow instrument identification without participation of human experts. Therefore, such systems can be valuable assistance for users of audio data searching for specific timbre, especially if they are not experienced musicians and when the amount of available audio data is huge, thus making manual searching impractical, if possible at all. When combined with a melody-searching system, automatic instrument classification may provide a handy tool for finding favorite tunes performed by favorite instruments in audio databases.

MAIN THRUST

Research on automatic classification of musical instruments so far has been performed mainly on isolated, singular sounds; works on polyphonic sounds usually aim at source separation and operations like pitch tracking of these sounds (Viste & Evangelista, 2003). Most commonly used data include MUMS compact discs (Opolko & Wapnick, 1987), the University of Iowa samples (Fritts, 1997), and IRCAM's Studio on Line (IRCAM, 2003).

A broad range of data mining techniques was applied in this research, aiming at extraction of information hidden in audio data (i.e., sound features that are common for a given instrument and differentiate it from the others). Descriptions of musical instrument sounds are usually subjective, and finding appropriate numeric descriptors (parameters) is a challenging task. Sound parameterization is arbitrarily chosen by the researchers, and the parameters may reflect features that are known to be important for the human in the instrument recognition task, like descriptors of sound evolution in time (i.e., onset features, depth of sound vibration, etc.), subjective timbre features (i.e., brightness of the sound), and so forth. Basically, parameters characterize coefficients of sound analysis, since they are relatively easy to calculate.

On the basis of parameterization, further research can be performed. Clustering applied to parameter vectors reveals similarity among sounds and adds a new glance on instrument classification, usually based on instrument construction or sound articulation. Decision rules and trees allow identification of the most descriptive sound features. Transformation of sound parameters may produce new descriptors better suited for automatic instrument classification. The classification can be performed hierarchically, taking instrument families and articulation into account. Classifiers represent a broad range of methods, from simple statistic tools to new advanced algorithms rooted in artificial intelligence.

Parameterization Methods

Sound is a physical disturbance in the medium (e.g., air) through which it is propagated (Whitaker & Benson, 2002). Periodic fluctuations are perceived as sound having pitch. The audible frequency range is about 20-20,000 Hz (hertz or cycles per second). The parameterization aims at capturing most distinctive sound features regarding sound amplitude evolution in time, static spectral features (frequency contents) of the most stable part of the sound, and evolution of frequency content in time. These features are based on the Fourier spectrum and time-frequency sound representations like wavelet transform. Some analyses are adjusted to the properties of human hearing, which perceives changes of sound amplitude and frequency in a logarithmic-like manner (e.g., frequency contents analysis in mel scale). The results based on such analysis are easier to interpret in subjective terms. Also, statistic and mathematic operations are applied to the sound representation, yielding good results, too. Some descriptors require calculating pitch of the sound, and any inaccuracies in pitch calculation (e.g., octave errors) may lead to erroneous results.

Parameter sets investigated in the research are usually a mixture of various types, since such combinations allow capturing more representative sound description for instrument classification purposes.

The following analysis and parameterization methods are used to describe musical instrument sounds:

- Autocorrelation and cross-correlation functions investigating periodicity of the signal and statistical parameters of spectrum obtained via Fourier transform: average amplitude and frequency variations (wide in vibrated sounds), standard deviations (Ando & Yamaguchi, 1993).
- Contents of selected groups of partials in the spectrum (Pollard & Jansson, 1982; Wieczorkowska, 1999a), including amount of even and odd harmonics (Martin & Kim, 1998), allowing identification of clarinet sounds.
- Vibrato strength and other changes of sound features in time (Martin & Kim, 1998; Wieczorkowska et al., 2003) and temporal envelope of the sound.
- Statistical moments of the time wave, spectral centroid (gravity center), coefficients of cepstrum (i.e., the Fourier transform applied to the logarithm of amplitude plot of the spectrum), constant-Q coefficients (i.e., for logarithmically-spaced spectral bins) (Brown, 1999; Brown et al., 2001).
- Wavelet analysis, providing time-frequency plot based on decomposition of sound signal into functions called wavelets (Kostek & Czyzewski, 2001; Wieczorkowska, 1999b).
- Mel-frequency coefficients (i.e., in mel scale, adjusted to the properties of human hearing) and linear prediction cepstral coefficients, where future values are estimated as a linear function of previous values (Eronen, 2001).
- Multidimensional Scaling Analysis (MSA) trajectories obtained through Principal Component Analysis (PCA) applied to the constant-Q spectral snapshots to determine the most significant attributes of each sound (Kaminskyj, 2002). PCA transforms a set of variables into a smaller set of