

# **Automatic Music Timbre Indexing**

**Xin Zhang\***

Department of Computer Science  
University of North Carolina at Charlotte  
9201 University City Boulevard  
Charlotte, NC 28211  
USA

voice: +1 704-687-8546  
email: xinzhang@uncc.edu

**Zbigniew W. Ras**

Department of Computer Science  
University of North Carolina at Charlotte  
9201 University City Boulevard  
Charlotte, NC 28211  
USA

Department of Intelligent Systems  
Polish-Japanese Institute of Information Technology  
Warsaw 02-008  
Poland

voice: +1 704-687-4567  
fax: +1 704-687-3516  
email: ras@uncc.edu

**(\* Corresponding author)**

# Automatic Music Timbre Indexing

Xin Zhang, University of North Carolina at Charlotte, USA

Zbigniew W. Ras, University of North Carolina at 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, while there are 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, a 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 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. 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.

Therefore, 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  $j^{\text{th}}$  harmonic peak of the  $k^{\text{th}}$  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.

## **Feature Extraction**

Methods in research on automatic musical instrument sound classification go back to last few years. So far, there is no standard parameterization used as a classification basis. The sound descriptors used are based on various methods of analysis in time domain, spectrum domain, time-frequency domain and cepstrum with Fourier Transform for spectral analysis being most common, such as Fast Fourier Transform, Short-time Fourier Transform, Discrete Fourier Transform, and so on. Also, wavelet analysis gains increasing interest for sound and especially for musical sound analysis and representation. Based on recent research performed in this area, MPEG proposed an MPEG-7 standard, in which it described a set of low-level sound temporal and spectral features. However, a sound segment of note played by a music instrument is known to have at least three states: transient state, quasi-steady state and decay state. Vibration pattern

in a transient state is known to significantly differ from the one in a quasi-steady state. Temporal features in differentiated states enable accurate instrument estimation.

These acoustic features can be categorized into two types in terms of size:

- Acoustical instantaneous features in time series: a huge matrix or vector, where data in each row describe a frame, such as Power Spectrum Flatness, and Harmonic Peaks, etc. The huge size of data in time series is not suitable for current classification algorithms and data mining approaches.
- Statistical summation of those acoustical features: a small vector or single value, upon which classical classifiers and analysis approaches can be applied, such as Tristimulus (Pollard and Jansson, 1982), Even/Odd Harmonics (Kostek and Wierzchowska, 1997), averaged harmonic parameters in differentiated time domain (Zhang and Ras, 2006A), etc.

### **Machine Learning Classifiers**

The classifiers, applied to the investigations on musical instrument recognition and speech recognition, represent practically all known methods: Bayesian Networks (Zweig, 1998; Livescu and Bilmes, 2003), Decision Tree (Quinlan, 1993; Wierzchowska, 1999), K-Nearest Neighbors algorithm (Fujinaga and McMillan 2000; Kaminskyj and Materka 1995), Locally Weighted Regression (Atkeson and Moore, 1997), Logistic Regression Model (le Cessie and Houwelingen, 1992), Neural Networks (Dziubinski, Dalka and Kostek 2005) and Hidden Markov Model (Gillet and Richard 2005), etc. Also, hierarchical classification structures have been widely used by researchers in this area (Martin and Kim, 1998; Eronen and Klapuri, 2000), where sounds have been first categorized to different instrument families (e.g. the String family,

the Woodwind Family, the Percussion Family, etc), and then been classified into individual categories (e.g. Violin, Cello, flute, etc.)

## **FUTURE TRENDS**

The classification performance relies on sound items of the training dataset and the multi-pitch detection algorithms. More new temporal features in time-variation against background noise and resonance need to be investigated. **Timbre detection** of sounds with overlapping in homogeneous pitches from different instruments can be a very interesting and challenging area.

## **CONCLUSION**

**Timbre detection** is one of the most important sub-tasks for content based indexing. In Automatic Music Timbre Indexing, timbre is estimated based on computation of the content of audio data in terms of acoustical features by **machine learning** classifiers. An automatic music timbre indexing system should have at least the following components: **sound separation**, **feature extraction**, and hierarchical timbre classification. We observed that sound separation based on multi-pitch trajectory significantly isolated heterogeneous harmonic sound sources in different pitches. Carefully designed temporal parameters in the differentiated time-frequency domain together with the **MPEG-7** low-level descriptors have been used to briefly represent subtle sound behaviors within the entire pitch range of a group of western orchestral instruments. The results of our study also showed that Bayesian Network had a significant better performance than Decision Tree, Locally Weighted Regression and Logistic Regression Model.

## **REFERENCES**

- Atkeson, C.G., Moore A.W., and Schaal, S. (1997). *Locally Weighted Learning for Control*, Artificial Intelligence Review. Feb. 11(1-5), 75-113.
- Balzano, G.J. (1986). *What are musical pitch and timbre?* Music Perception - an interdisciplinary Journal. 3, 297-314.
- Bregman, A.S. (1990). *Auditory scene analysis, the perceptual organization of sound*, MIT Press
- Cadoz, C. (1985). *Timbre et causalite*, Unpublished paper, Seminar on Timbre, Institute de Recherche et Coordination Acoustique / Musique, Paris, France, April 13-17.
- Dziubinski, M., Dalka, P. and Kostek, B. (2005) *Estimation of Musical Sound Separation Algorithm Effectiveness Employing Neural Networks*, Journal of Intelligent Information Systems, 24(2/3), 133–158.
- Eronen, A. and Klapuri, A. (2000). *Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features*. In proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, Plymouth, MA, 753-756.
- Fujinaga, I., McMillan, K. (2000) *Real time Recognition of Orchestral Instruments*, International Computer Music Conference, 141-143.
- Gillet, O. and Richard, G. (2005) *Drum Loops Retrieval from Spoken Queries*, Journal of Intelligent Information Systems, 24(2/3), 159-177
- Herrera. P., Peeters, G., Dubnov, S. (2003) *Automatic Classification of Musical Instrument Sounds*, Journal of New Music Research, 32(19), 3–21.
- Kaminskyj, I., Materka, A. (1995) *Automatic source identification of monophonic musical instrument sounds*, the IEEE International Conference On Neural Networks.

- Kostek, B. and Wieczorkowska, A. (1997). *Parametric Representation of Musical Sounds*, Archive of Acoustics, Institute of Fundamental Technological Research, Warsaw, Poland, 22(1), 3-26.
- le Cessie, S. and van Houwelingen, J.C. (1992). *Ridge Estimators in Logistic Regression*, Applied Statistics, 41, (1 ), 191-201.
- Livescu, K., Glass, J., and Bilmes, J. (2003). *Hidden Feature Models for Speech Recognition Using Dynamic Bayesian Networks*, in Proc. Euro-speech, Geneva, Switzerland, September, 2529-2532.
- Martin, K.D., and Kim, Y.E. (1998). *Musical Instrument Identification: A Pattern-Recognition Approach*. 136th Meeting of the Acoustical Soc. of America, Norfolk, VA. 2pMU9.
- Pollard, H.F. and Jansson, E.V. (1982). *A tristimulus Method for the spectification of Musical Timbre*. Acustica, 51, 162-171
- Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.
- Wieczorkowska, A. (1999). *Classification of musical instrument sounds using decision trees*, in the 8th International Symposium on Sound Engineering and Mastering, ISSEM'99, 225-230.
- Wieczorkowska, A., Wroblewski, J., Synak, P., and Slezak, D. (2003). *Application of Temporal Descriptors to Musical Instrument Sound*, *Journal of Intelligent Information Systems, Integrating Artificial Intelligence and Database Technologies*, July, 21(1), 71-93.
- Zhang, X. and Ras, Z.W. (2006A). *Differentiated Harmonic Feature Analysis on Music Information Retrieval For Instrument Recognition*, proceeding of IEEE International Conference on Granular Computing, May 10-12, Atlanta, Georgia, 578-581.

Zhang, X. and Ras, Z.W. (2006B). *Sound Isolation by Harmonic Peak Partition For Music Instrument Recognition*, Fundamenta Informaticae Journal Special issue on Tilings and Cellular Automata, IOS Press, 2006

Zweig, G. (1998). *Speech Recognition with Dynamic Bayesian Networks*, Ph.D. dissertation, Univ. of California, Berkeley, California.

ISO/IEC JTC1/SC29/WG11 (2002). MPEG-7 Overview. Available at <http://mpeg.telecomitalia.com/standards/mpeg-7/mpeg-7.htm>

## **KEY TERMS AND THEIR DEFINITIONS.**

**Automatic Indexing:** Automatically identifies precise, relevant, clips of content within audio sources.

**Sound Separation:** The process of isolating sound sources within a piece of sound.

**Feature Extraction:** The process of generating a set of descriptors or characteristic attributes from a binary musical file.

**Harmonic:** A set of component pitches in mathematical relationship with the fundamental frequency.

**Hierarchical Classification:** Classification in a top-down order. First identify musical instrument family types, and then categorize individual or groups of instruments within the instrument family.

**Machine Learning:** A study of computer algorithms that improve their performance automatically based on previous results.

**MPEG-7:** A Multimedia Content Description Interface standardizes descriptions for audio-visual content by Moving Picture Experts Group.

**Quasi-steady State:** A steady state where frequencies are in periodical patterns.

**Short-Time Fourier Transform:** By using an analysis window, e.g. a hamming window, signal is evaluated with elementary functions that are localized in time and frequency domains simultaneously.

**Timbre:** Describes those characteristics of sound, which allow the ear to distinguish one instrument from another.

**Time-Frequency Domain:** A time series of analysis windows, where patterns are described in frequency domain.