MIRAI: Multi-hierarchical Music Automatic Indexing and Retrieval System

Zbigniew W. Raś[1][2], **Xin Zhang**[1]

Abstract: Recently, numerous successful approaches have been developed for instrument recognition in monophonic sounds. Unfortunately, none of them can be successfully applied to polyphonic sounds. Identification of music instruments in polyphonic sounds is still difficult and challenging. This has stimulated a number of research projects on music sound separation and new features development for content-based automatic music information retrieval. The paper introduces several temporal features based on pitch to improve automatic music instrument recognition. The results from experiments show that these new features, with the pitch information removed from them, tend to provide less distraction for timber estimation. Sometime, the addition of new features to the database of music instruments does not help and related classifiers still do not perform well. One possibility to handle this problem is to build classifiers which learn not only the descriptions of music instruments but also their generalizations on different granularity levels. We show that by introducing several optional hierarchical classifications of musical instruments and constructing related classifiers, we increase a chance to build a system of good performance in terms of successful indexing of music by instruments and their types.

Keywords: music information retrieval, automatic indexing, knowledge discovery.

1. Introduction

The ultimate goal of our project is a creation of a web-based storage and retrieval system, called *MIRAI*, which can automatically index musical input (of polyphonic type) into FS-tree type database and answer queries requesting specific musical pieces [http://www.mir.uncc.edu/]. When *MIRAI* receives a musical waveform, it divides that waveform into segments of equal size and then its classifiers identify the most dominating musical instruments and emotions associated with that segment. Our database of musical instrument sounds is constantly growing and it currently has about 4,000 sound objects and more than 1,100 features. Each sound object is represented as a temporal sequence of approximately 150-300 tuples which gives us a temporal database of more than 1,000,000 tuples, each one represented as a vector of about 1,100 features. This database is mainly used to learn classifiers for automatic indexing of musical instrument sounds. A separate database containing

[1] Department of Computer Science, University of North Carolina, Charlotte, N.C. 28223, USA
e-mail: `ras@uncc.edu, cynthiaxz@uncc.edu`

[2] Polish-Japanese Institute of Information Technology, ul. Koszykowa 86, 02-008 Warszawa
e-mail: `ras@pjwstk.edu.pl`

longer musical pieces which are indexed according to their scalar relations is used for automatic indexing of emotions. Both databases have to be semantically reach enough (in terms of successful sound separation and recognition) so the constructed classifiers have a high level of accuracy in recognizing musical instruments and/or their types when music is polyphonic. This paper shows that by adding new temporal non-MPEG7 features to our database of musical instrument sounds, we can improve the confidence of *MIRAI* classifiers. The same, the precision of *MIRAI* retrieval engine is also getting improved.

Recently, a number of acoustical features for the construction of a computational model for music timbre estimation have been investigated in Music Information Retrieval (MIR) area. 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. It is rather a subjective quality, defined by ANSI as the attribute of auditory sensation, in terms of which a listener can judge that two sounds, similarly presented and having the same loudness and pitch, are different. Such definition is clearly subjective and not of much use for automatic sound timbre classification. Therefore, musical sounds must be very carefully parameterized to allow automatic timbre recognition. The real use of timbre-based grouping of music and dimensional approach to timbre description is very nicely discussed in (Bregman, 1990). 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 (FFT), Short-Time Fourier Transform (STFT), Discrete Fourier Transform (DFT), 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 [http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm], in which a set of low-level sound temporal and spectral features is described.

A short digital musical file may consist of a huge number of integers in its content to represent sound vibration in time. 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 integers in the time-order sequence will be 2,646,000, which makes it a very big data item. Therefore, features to capture subtle changes are normally in a form of matrix or vector. High dimensionality brings another challenge to MIR area. Researchers explored different statistical summations to describe signatures of music instruments based on vectors or matrices in features, such as Tristimulus parameters (Pollard and Jansson, 1982), brightness (Fujinaga and McMillan, 2000), and irregularity (Wold et al., 1996), etc. Flattening these features for traditional classifiers increases the number of features. Authors in their previous work already evaluated the quality of classifiers based on additional features that they developed and added to MPEG7 features and other popular features, bringing the total number of features to three hundred. In this paper, authors focus on development of new features to improve the classification efficiency for harmonic sounds, which have steady pitch state.

Authors also show that by introducing several optional hierarchical classifications of musical instruments and constructing related classifiers, we increase a chance to build a system of good performance in terms of successful indexing of music by instruments and their types.

2. Audio Features

There are many ways to present audio features by different categorization method. In our system, audio features are first categorized as MPEG7 descriptors and other/non-MPEG7 descriptors in the acoustical perspective of view, where both spectrum features and temporal features are included. Then, the new temporal features are presented. Finally, a derivative database of those features with single valued data for KDD classification is demonstrated. The spectrum features have two different frequency domains: Hz frequency and Mel frequency. Frame size is carefully designed to be 120ms, so that the 0th octave G (the lowest pitch in our audio database) can be detected. The hop size is 40ms with a overlapping of 80ms. Since the sample frequency of all the music objects is 44,100Hz, the frame size is 5292. A hamming window is applied to all STFT transforms to avoid jittering in the spectrum.

3. MPEG7 Based Descriptors

Based on latest research in the area, MPEG published a standard group of features for the digital audio content data. They are either in the frequency domain or in the time domain. A STFT with hamming window has been applied to the sample data, where each frame generates a set of instantaneous values.

Spectrum Centroid describes the center-of-gravity of a log-frequency power spectrum in the following formulas. It economically indicates the pre-dominant frequency range. $P_x(k)$ is a power spectrum coefficient. Coefficients under 62.5Hz have been grouped together for fast computation.

1.) $P_x(k)$, $k = 0, ..., \frac{NFFT}{2}$
2.) $C = \sum_n \log_2 \big(f(n)/1000\big) P_x^{'}(n) / \sum_n P_x^{'}(n)$.

A mean value and standard deviation of all frames have been used to describe the Spectrum Centroid of a music object.

Spectrum Spread is the Root of Mean Square value of the deviation of the Log frequency power spectrum with respect to the gravity center in a frame. Like Spectrum Centroid, it is an economic way to describe the shape of the power spectrum.

3.) $S = \sqrt{\sum_n \big(\big(\log_2(f(n)/1000\big) - C)^2 P'(n)\big) / \sum_n P'(n)}$.

A mean value and standard deviation of all frames have been used to describe the Spectrum Spread of a music object.

Spectrum Flatness describes the flatness property of the power spectrum within a frequency bin. The value of each bin is treated as an attribute value in the database.

Spectrum Basis Functions are used to reduce the dimensionality by projecting the spectrum from high dimensional space to low dimensional space with compact salient statistical information.

Harmonic Centroid is computed as the average, over the sound segment duration, of the instantaneous Harmonic Centroid within a frame. The instantaneous Harmonic Spectral Centroid is computed as the amplitude in linear scale weighted mean of the harmonic peak of the spectrum.

Harmonic Spread is computed as the average over the sound segment duration of the instantaneous harmonic spectral spread of frame. The instantaneous harmonic spectral spread is computed as the amplitude weighted standard deviation of the harmonic peaks of the spectrum with respect of the instantaneous harmonic spectral centroid.

Harmonic Variation is defined as the mean value over the sound segment duration of the instantaneous harmonic spectral variation. The instantaneous harmonic spectral variation is defined as the normalized correlation between the amplitude of the harmonic peaks of two adjacent frames.

Harmonic Deviation is computed as the average over the sound segment duration of the instantaneous Harmonic Spectral Deviation in each frame. The instantaneous Harmonic Spectral Deviation is computed as the spectral deviation of the log amplitude components from a global spectral envelope.

Log Attack Time is defined as the logarithm of the time duration between the time when the signal starts to the time it reaches its stable part, where the signal envelope is estimated by computing the local mean square value of the signal amplitude in each frame.

$$LAT = log_{10}(T1 - T0)$$

where $T0$ is the time when the signal starts, $T1$ is the time the signal reaches its sustained part of maximum part.

Harmonicity Rate is the proportion of harmonics in the power spectrum. It describes the degree of harmonicity of a frame. It is computed by the normalized correlation between the signal and a lagged representation of the signal.

Upper Limit of Harmonicity describes the frequency beyond which the spectrum cannot be considered harmonic. It is calculated based on the power spectrum of the original and a comb-filtered signal.

Spectral Centroid is computed as the power weighted average of the frequency bins in the power spectrum of all the frames in a sound segment with a Welch method.

Temporal Centroid is calculated as the time average over the energy envelope.

4. Temporal Features Based on Pitch

Pitch trajectories of instruments behave very differently in time. Authors designed parameters to capture the power change in time.

Pitch Trajectory Centroid PC is used to describe the center of gravity of the power of the fundamental frequency during the quasi-steady state.

$$PC = \frac{\sum_{n=1}^{lenght(P)} [\frac{n}{length(P)} P(n)]}{\sum_{n=1}^{lenght(P)} P(n)}$$

where P is the pitch trajectory in the quasi-steady state, n is the n^{th} frame.

Pitch Trajectory Spread PS is the RMS deviation of the Pitch Trajectory with respect to its gravity center.

$$PS = \sqrt{\frac{\sum_{n=1}^{length(P)} [\frac{n}{length(P)} - PC]^2 \cdot P(n)}{\sum_{n=1}^{length(P)} P(N)}}.$$

Pitch Trajectory Max Angle PM is an angle of the normalized power maximum vs. its normalized frame position along the trajectory in the quasi-steady state.

$$PM = \frac{\left[\frac{MAX(P(n)) - P(0)}{\frac{1}{length(P)} \cdot \sum_{n=1}^{length(P)} P(n)}\right]}{\frac{F(n) - F(0)}{length(P)}},$$

where $F(n)$ is the position of n^{th} frame in the steady state.

Harmonic Relation is a vector to describe the relationship among the harmonic partials.

$$HR = \frac{1}{m} \sum_{j=1}^{m} \frac{H_j}{H_0},$$

where m is the total number of frames in the steady state, H_j is the j^{th} harmonic peak.

5. Other Descriptors

In order to obtain compact representation of musical acoustical features, the following descriptors have been used in the paper.

Fundamental Frequency is the frequency that best explains the periodicity of a signal. The ANSI definition of psycho-acoustical terminology says that *pitch is the auditory attribute of a sound according to which sounds can be ordered on a scale from low to high.* It is estimated by the maximum likelihood of candidate frequencies (Zhang, Marasek, Raś, 2007).

Vector Descriptors. Since a value of a descriptor is a matrix, statistical value retrieval has been performed for traditional classifiers. These statistical values are maximum, minimum, mean value, and the standard deviation of the matrix, maximum, minimum, mean value of dissimilarity of each column.

Tristimulus parameters and similar parameters describe the ratio of the amplitude of a harmonic partial to the total harmonic partials (Pollard and Jansson, 1982). They are: first modified tristimulus parameter, power difference of the first and the second tristimulus parameter, grouped tristimulus of other harmonic partials, odd and even tristimulus parameters.

Brightness is calculated as the proportion of the weighted harmonic partials to the harmonic spectrum.

Transient, steady and decay duration. The transient duration is considered as the time to reach the quasi-steady state of fundamental frequency (Zhang and Raś, 2006). In this duration the sound contains more timbre information than pitch information that is highly relevant to the fundamental frequency. Thus differentiated harmonic descriptors values in time are calculated based on the subtle change of the fundamental frequency. The duration after the quasi-steady state is treated as the decay state. All the duration values are normalized by the length of their corresponding audio objects.

Zero crossing counts the number of times that the signal sample data changes signs in a frame (Tzanetakis, Cook, 2002).

Spectrum Centroid describes the gravity center of the spectrum (Wieczorkowska et al., 2003).

Roll-off is a measure of spectral shape, which is used to distinguish between voiced and unvoiced speech. The roll-off is defined as the frequency below which C percentage of the accumulated magnitudes of the spectrum is concentrated, where C is an empirical coefficient.

$$\sum_{k=1}^{K} |X_i(k)| \leq C \cdot \sum_{k=1}^{K} |X_i(k)|.$$

Flux is used to describe the spectral rate of change. It is computed by the total difference between the magnitude of the FFT points in a frame and its successive frame.

$$F_i = \sum_{k=1}^{\frac{N}{2}} (|X_i(k)| - |X_{i-1}(k)|)^2.$$

Mel frequency cepstral coefficients describe the spectrum according to the human perception system in the Mel scale. They are computed by grouping the STFT points of each frame into a set of 40 coefficients by a set of 40 weighting curves with logarithmic transform and a discrete cosine transform (DCT).

6. Hierarchical Classifications of Musical Instruments

There are many ways to categorize music instruments, such as by playing methods, by instrument type, or by other generalization concepts. Any categorization process can be represented as a hierarchical schema which is used by a cooperative query answering system to handle failing queries. By definition, a cooperative system is relaxing a failing query with a goal to find its smallest generalization which will not fail. Two different hierarchical schemas (Ras et al., 2007), used as models of a decision attribute, have been investigated in authors previous research: Hornbostel-Sachs classification of musical instruments and classification of musical instruments by articulation, with 15 different articulation methods (seen as attribute values): blown, bowed, bowed vibrato, concussive, hammered, lip-vibrated, martele, muted, muted vibrato, percussive, picked, pizzicato, rubbed, scraped and shaken. Each hierarchical classification represents a unique decision attribute, in a database of music instruments, leading to a construction of a new classifier and the same to a different system for automatic indexing of music by instruments and their types.

The main classification is based on the Hornbostel and Sachs system (with extensions)(Hornbostel, 1914). Basic classification includes aerophones (wind instruments), chordophones (string instruments), idiophones (made of solid, non-stretchable, resonant material), and membranophones (mainly drums); idiophones and membranophones are together classified as percussion. Additional groups include electrophones, i.e. instruments where the acoustical vibrations are produced by electric or electronic means (electric guitars, keyboards, synthesizers), complex mechanical instruments (including pianos, organs, and other mechanical music makers), and special instruments (include bullroarers, but they can be classified as free aerophones). Each category can be further subdivided into groups, subgroups etc. and finally into instruments. Aerophones subcategories are also called woodwinds or brass, but this criterion is not based on the material the instrument is made of, but rather on the method of sound production. In woodwinds, the change of pitch is mainly obtained by the change of the length of the column of the vibrating air. Additionally, over-blow is applied to obtain second, third or fourth harmonic to become the fundamental. In brass instruments, over-blows are very easy because of wide bell, and therefore they are seen as the main method of pitch changing.

Sounds can be classified according to the articulation which can be performed in three ways: (1) sustained or non-sustained sounds, (2) muted or not muted sounds, (3) vibrated and not vibrated sounds. This partition may be difficult to obtain, since vibration does not have to appear in the entire sound; some changes may be visible, but no clear vibration. Also, brass is sometimes played with moving the mute in and out of the bell.

According to the contents of the spectrum, the musical instrument sounds can be classified into the following three types: (1) harmonic spectrum, (2) continuous spectrum, or (3) mixed spectrum. Most of music instrument sounds of definite pitch have some noises/continuity in their spectra. According to MPEG-7 classification [4], there are four classes of musical instrument sounds: (1) Harmonic, sustained, coherent sounds - well detailed in MPEG-7, (2) Nonharmonic, sustained, coherent sounds, (3) Percussive, nonsustained sounds - well detailed in MPEG-7, (4) Noncoherent, sustained sounds. This also can be misleading, since pizzicato is not clearly present in this classification, as harmonic, non-sustained sound.

We can also cluster musical instruments in many other ways and the same generate many possible hierarchical structures each defining a new decision-attribute.

A formal framework for evaluation and comparison of different classifications of musical sounds is discussed in the rest of this section. Musical instruments are represented as leaves of a hierarchical decision attribute, denoted in our case by d and its different types and subtypes are represented as internal nodes of d. In our database called *MIRAI*, musical instruments are represented as sample musical sounds described by a large number sound features, denoted by A (Ras et al., 2007). The goal of each classification is to find descriptions of musical instruments or their classes (values of attribute d) in terms of values of attributes from A. Each classification results in a classifier which can be evaluated using standard methods like bootstrap or cross-validation. In our research we use ten-fold cross-validation.

Let us assume that $S = (X, A \cup \{d\}, V)$ is a decision system, where d is a hierarchical attribute. We also assume that $d_{[i_1,...,i_k]}$ (where $1 \leq i_j \leq m_j$, $j = 1, 2..., k$) is a child of $d_{[i_1,...,i_{k-1}]}$ for any $1 \leq i_k \leq m_k$. Clearly, attribute d has $\Sigma\{m_1 \cdot m_2 \cdot ... \cdot m_j : 1 \leq j \leq k\}$ values, where $m_1 \cdot m_2 \cdot ... \cdot m_j$ shows the upper bound for the number of values at the level j of d. By $p([i_1, ..., i_k])$ we denote a path (d, $d_{[i_1]}$, $d_{[i_1,i_2]}$, $d_{[i_1,i_2,i_3]}$,..., $d_{[i_1,...,i_{k-1}]}$, $d_{[i_1,...,i_k]}$) leading from the root of the hierarchical attribute d to its descendant $d_{[i_1,...,i_k]}$.

Let us assume that R_j is a set of classification rules extracted from S, representing a part of a rule-based classifier $R = \bigcup\{R_j : 1 \leq j \leq k\}$, and describing all values of d at level j. The quality of a classifier at level j of attribute d can be checked by calculating $Q(R_j) = \frac{\sum\{sup(r)\cdot conf(r):r\in R_j\}}{\sum\{sup(r:r\in R_j\}}$, where $sup(r)$ is the support of the rule r in S and $conf(r)$ is its confidence. Then, the quality of the rule-based classifier R can be checked by calculating $Q(\bigcup\{R_j : 1 \leq j \leq k\}) = \frac{\sum\{Q(R_j):1\leq j\leq k\}}{k}$.

The quality of a tree-based classifier can be given by calculating its quality for every node of a hierarchical decision attribute d. Let us take a node $d_{[i_1,...,i_k]}$ and the path $p([i_1,...,i_k])$ leading to that node from the root of d. There is a set of classification rules $R_{[i_1,...,i_m]}$, uniquely defined by the tree-based classifier, assigned to a node $d_{[i_1,...,i_m]}$ of a path $p([i_1,...,i_k])$, for every $1 \leq m \leq k$. Now, we define $Q(R_{[i_1,...,i_m]})$ as $\frac{\sum\{sup(r)\cdot conf(r):r\in R_{[i_1,...,i_m]}\}}{\sum\{sup(r):r\in R_{[i_1,...,i_m]}\}}$. Then, the quality of a tree-based classifier for a node $d_{[i_1,...,i_m]}$ of the decision attribute d can be checked by calculating $Q(d_{[i_1,...,i_m]}) = \prod\{Q(R_{[i_1,...,i_j]}) : 1 \leq j \leq m\}$. In our experiments, presented in Section 4 of this paper, we use *J48 Tree* as the tool to build tree-based classifiers. Also, their performance on level m of the attribute d is checked by calculating $Q(d_{[i_1,...,i_m]})$ for every node $d_{[i_1,...,i_m]}$ at the level m. Finally, the performance of both classifiers is checked by calculating $Q(\bigcup\{R_j : 1 \leq j \leq k\})$ (the first method we proposed).

Learning values of a decision attribute at different generalization levels is extremely important not only for designing and developing an automatic indexing system of possibly highest confidence but also for handling failing queries. Values of a decision attribute and their generalizations are used to construct atomic queries of a query language built for retrieving musical objects from *MIRAI* Database (see http://www.mir.uncc.edu). When query fails, the cooperative strategy (Gaasterland, 1997), (Godfrey, 1993) will try to find its smallest generalization which does not fail. Clearly, by having a variety of different hierarchical structures available for d we have better chance not only to succeed but succeed with a more optimal generalization of an instrument class.

7. Classification

The classifiers, applied in the investigations on musical instrument recognition, represent practically all known methods. The authors applied Decision Tree-J48 in the classification. Decision Tree-J48 is a supervised classification algorithm, which has been extensively used for machine learning and pattern recognition. A Tree-J48 is normally constructed top-down, where parent nodes represent conditional attributes and leaf nodes represent decision outcomes. It first chooses a most informative attribute that can best differentiate the dataset; it then creates branches for each interval of the attribute where instances are divided into groups; it repeats creating sub-branches until instances are clearly separated in terms of the decision attribute; finally it tests the tree by new instances in a test dataset.

8. Experiments

We used a database of 1569 music recording sound objects from McGill University Master Samples CD Collection instruments, which has been widely used for research on musical instrument recognition all over the world. All classifiers were 10-fold cross validation with a split of 90% training and 10% testing. We used WEKA for all classifications and Rough Set Library for data reduction.

Classification has been performed on different levels in a music instrument categorization schema. The first level instrument types are aerophone, chordophone, idiophone; the second level types include aero free, aero free-reed, aero lip-vibrated, aero side, aero single-reed, chrd composite, chrd simple, idio concussion, idio rubbed, idio scraped, idio shaken, idio struck; the third level types are instruments (e.g., violin, piano, etc.).

$J48 - Tree$	$with.new.Fe$	$with.new.Fe$	$without.new.Fe$	$without.new.Fe$
	A	B	A	B
Idiophone	91.8 %	94.9 %	91.4 %	89.8 %
Chordophone	92.7 %	89.1 %	89.5 %	86.7 %
Aerophone	90.8 %	92.9 %	87.9 %	91.3 %
Overall	91.7 %		89.2 %	

Table 1. Results of Classification in 1^{th} level

$J48 - Tree$	$with.new.Fe$	$with.new.Fe$	$without.new.Fe$	$without.new.Fe$
	A	B	A	B
Lip	84.6 %	83.5 %	78.7 %	74.7 %
Side	68.4 %	68.4 %	66.7 %	63.2 %
Reed	73.5 %	86.2 %	66.7 %	82.8 %
Composite	90.0 %	84.1 %	85.6 %	88.0 %
Simple	78.3 %	85.7 %	77.3 %	81.0 %
Rubbed	90.9 %	100.0 %	100.0 %	40.0 %
Shaken	82.6 %	95.0 %	76.2 %	80.0 %
Struck	92.3 %	82.8 %	81.5 %	75.9 %
Overall	84.4 %		79.3 %	

Table 2. Results of Classification in 2^{th} level

Table2 and Table3 show that new features significantly improved the classification of individual instruments.

9. Conclusions

The results from experiments show that the new features, with the pitch information removed from them, tend to provide less distraction for timber estimation. The *pitch-removed* features significantly improved the classification of individual

$J48 - Tree$	$with.new.Fe$	$with.new.Fe$	$without.new.Fe$	$without.new.Fe$
	A	B	A	B
Con-clarinet	100.0 %	60.0 %	83.3 %	100.0 %
Electric-bas	100.0 %	73.3 %	93.3 %	93.3 %
Flute	100.0 %	50.0 %	60.0 %	75.0 %
Steel-drums	100.0 %	66.7 %	50.0 %	66.7 %
Tuba	100.0 %	100.0 %	100.0 %	85.7 %
Vibraphone	87.5 %	93.3 %	78.6 %	73.3 %
Cello	87.0 %	95.2 %	86.7 %	61.9 %
Violin	84.0 %	77.8 %	66.7 %	59.3 %
Piccolo	83.3 %	50.0 %	60.0 %	60.0 %
Marimba	82.4 %	87.5 %	83.3 %	93.8 %
C-trumpet	81.3 %	76.5 %	87.5 %	82.4 %
Alto-flute	80.0 %	80.0 %	80.0 %	80.0 %
English-horn	80.0 %	57.1 %	42.9 %	42.9 %
Trombone	80.0 %	94.1 %	81.3 %	76.5 %
Piano	79.2 %	90.5 %	70.4 %	90.5 %
Double-bass	77.8 %	63.6 %	41.7 %	45.5 %
French-horn	76.9 %	76.9 %	71.4 %	76.9 %
Oboe	75.0 %	85.7 %	77.8 %	100.0 %
Electric-guitar	70.6 %	66.7 %	90.0 %	50.0 %
Saxophone	66.7 %	80.0 %	66.7 %	80.0 %
Viola	66.7 %	42.9 %	38.9 %	50.0 %

Table 3. Results of Classification in 3^{th} level

instruments. However, the higher level the classification is in, the less significant is the improvement provided by the new features. This may be caused by feature distraction or the confliction of the music schema, which future research will investigate.

10. Acknowledgment

This work is supported by the National Science Foundation under grant IIS-0414815.

References

Bregman, A.S. (1990) Auditory Scene Analysis, the Perceptual Organization of Sound. MIT Press.

Fujinaga, I., McMillan, K. (2000) Real Time Recognition of Orchestral Instruments, *I*nternational Computer Music Conference, 141–143

Gaasterland, T. (1997) Cooperative Answering Through Controlled Query Relaxation. *I*EEE Expert, Vol. 12, No. 5, 48–59

GODFREY, P. (1993) Minimization in Cooperative Response to Failing Database Queries, *I*nternational Journal of Cooperative Information Systems, Vol. 6, No. 2, 95–149

HORNBOSTEL, E. M. V., SACHS, C. (1914) Systematik der Musikinstrumente. Ein Versuch, *Z*eitschrift fur Ethnologie, Vol. 46, No. 4-5, 553-90, available at http://www.uni-bamberg.de/ppp/ethnomusikologie/HS-Systematik/HS-Systematik

LEWIS, R., ZHANG, X., RAŚ, Z. (2006) Blind Signal Separation of Similar Pitches and Instruments in a Noisy Polyphonic Domain. *Foundations of Intelligent Systems, 16th International Symposium, ISMIS 2006, Bari, Italy, September 2006, Proceedings. LNAI 4203, Subseries of Lecture Notes in Computer Science, Springer*, 228–237

POLLARD, H.F., JANSSON, E.V. (1982) A Tristimulus Method for the Specification of Musical Timbre, *Acustica, 51*, 162–171

RAS, Z. W., ZHANG, X., LEWIS, R. (2007) MIRAI: Multi-hierarchical, FS-Tree Based Music Information Retrieval System. *Rough Sets and Intelligent Systems Paradigms, International Conference, RSEISP 2007, Warsaw, Poland, June 28-30, 2007, Proceedings. LNAI 4585, Subseries of Lecture Notes in Computer Science, Springer*, 80–89

TZANETAKIS, G., COOK, P. (2002) Musical Genre Classification of Audio Signals, in *IEEE Trans. Speech and Audio Processing, July, Vol. 10*, 293–302

WIECZORKOWSKA, A., RAŚ, Z.W., ZHANG, X., LEWIS, R. (2007) Multi-way Hierarchic Classification of Musical Instrument Sounds. *Proceedings of the International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), IEEE Computer Society, April 26-28, 2007, in Seoul, South Korea*, 897–902

WIECZORKOWSKA, A., WRÓBLEWSKI, J., SYNAK, P., AND SLEZAK, D. (2003) Application of Temporal Descriptors to Musical Instrument Sound, in *Journal of Intelligent Information Systems, Springer, July, 21(1)*, 71–93

WOLD, E., BLUM, T., KEISLAR, D., AND WHEATON, J. (1996) Content-Based Classification, Search and Retrieval of Audio. *IEEE Multimedia, Fall*, 27–36

ZHANG, X., MARASEK, K., RAŚ, Z.W. (2007) Maximum Likelihood Study for Sound Pattern Separation and Recognition, in *Proceedings of the International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), IEEE Computer Society, April 26-28, 2007, in Seoul, South Korea*, 807–812

ZHANG, X. AND RAŚ, Z.W. (2006) Differentiated Harmonic Feature Analysis on Music Information Retrieval for Instrument Recognition, in *Proceeding of IEEE International Conference on Granular Computing, May 10-12, Atlanta, Georgia*, 578–581

Zhang, X., Raś, Z. W. (2007) Analysis of Sound Features for Music Timbre Recognition. *International Conference on Multimedia and Ubiquitous Engineering MUE 2007, 26-28 April 2007, Seoul, Korea. Edited by S. Kim, J. H. Park, N. Pissinou, T. Kim, W. C. Fang, D. Slezak, H. Arabnia, D. Howard. IEEE Computer Society*, 3–8