

# Multiple classifiers for different features in timbre estimation

Wenxin Jiang<sup>1</sup>, Xin Zhang<sup>3</sup>, Amanda Cohen<sup>1</sup>, Zbigniew W. Ras<sup>1,2</sup>

<sup>1</sup>Computer Science Department, University of North Carolina, Charlotte, N.C., USA,

<sup>2</sup>Polish-Japanese Institute of Information Technology, 02-008 Warsaw, Poland,

<sup>3</sup>Mathematics and Computer Science Department, University of North Carolina, Pembroke, N.C., USA

{wjiang3, xinzhang, acohen24, ras}@[uncc.edu](mailto:uncc.edu)

**Abstract** Computer storage and network techniques have brought a tremendous need to find a way to automatically index digital music recordings. In this paper, state of art acoustic features for timbre automatic indexing were explored to construct efficient classification models, such as decision tree as well as KNN. The authors built a database containing more than one million music instrument sound slices, each described by a large number of features including standard MPEG7 audio descriptors, features for speech recognition, and many new audio features developed by the authors, spanning from temporal space to spectral domain. Each classification model was tuned with feature selection based on its distinct characteristics for the blind sound separation system. Based on the experiment results, the authors proposed a new framework for MIR with multiple classifiers trained on different features. Inspired by the human recognition experience, timbre estimation based on the hierarchical structure of musical instrument families was investigated. A framework for the Cascade Classification System was proposed. The authors also discussed the issue of features and classifiers selection during the cascade classification process.

## Introduction

Automatic indexing of timbre is one of the main tasks in Music Information Retrieval in digital recordings. The use of timbre-based grouping of music is very nicely discussed in [5]. The classifiers, applied in investigations on musical instrument sound classification represent most of the known methods. One of the most popular classifiers is k-Nearest Neighbor (KNN) [9]. Other classifiers include Bayes decision rules, Gaussian mixture model [4], artificial neural networks [12], decision trees and rough set based algorithms [25], Hidden Markov Models (HMM), support vector machines (SVM) and other. However, the results for more than 10 instruments, explored in full musical scale range, generally are below 80%. Extensive review of parameterization and classification methods applied in research on this topic, with obtained results, is given in [13].

Typically a digital music recording, in the form of a binary file, contains a header and a body. The header stores file information such as length, number of channels, sampling rate, etc. Unless it is manually labeled, a digital audio recording has no description of timbre or other perceptual properties. It is a highly difficult 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 sampling 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 large data item.

Since these objects are not in a well-structured form with semantic meaning, this type of data is not suitable for most traditional data mining algorithms. Therefore, a number of features have been explored to give a higher-level representation of digital musical objects with structured and meaningful attributes based on acoustical expertise. Then these feature datasets can be intuitively used as system semantics, since they are computational and “known” to the computer system.

## *Pitch, melody and rhythm*

Pitch is the perceived quality of how high or low a sound is. This is chiefly a function of the fundamental frequency of the sound. In general pitch is regarded as becoming higher with increasing frequency and lower with decreasing frequency. The difference between two pitches is called an interval. A melody often consists of a sequence of pitches. The harmony, a musical line which adds support and dimension to the melody, can also consist of a sequence of pitches but is typically made up of a set of intervals, also known as chords.

There is another facet of music information which is called the temporal facet. It is the duration of musical events. Features such as tempo indicators and meter describe the rhythmic characteristics of an entire piece of music, although any of these features can be changed partway through a piece as is fitting. The tempo describes the overall speed at which a piece is to be played. Meter describes how many beats are in a measure which contributes to the overall rhythmic feel of the song. For example, a waltz typically has three beats in a measure, while a march may have either two or four beats in a measure. Other features like pitch duration, harmonic duration, and accents describe the rhythmic characteristics of specific notes. Those temporal events make up the rhythmic component of a musical work.

In music information retrieval area, a lot of research has been conducted on melody or rhythm matching based on pitch identification, which usually involves fundamental frequency detection. Utrecht University provides an overview of content-based Music Information Retrieval systems [1]. Around 43 MIR systems are listed; most of them are query by whistling/humming systems for melody retrieval. So far no system exists that can retrieve information about timbre in the literature or market, which indicates that it is an unsolved task.

## *Timbre*

According to the definition of American Standards Association, timbre is the quality of sound that is not loudness and pitch. It distinguishes different musical instruments playing the same note with identical pitch and loudness. So it is the most important and relevant facet of music information. People discern timbres from speech and music in everyday life.

Musical instruments usually produce sound waves with multiple frequencies. The frequencies are called harmonics, or harmonic partials. The lowest frequency is fundamental frequency  $f_0$ , which has an intimate relation with pitch. The remaining higher frequencies are called overtones. Along with the fundamental frequency, these harmonic partials make up the timbre, which is also called tone color. The aural distinction between different musical instruments is caused by the differences in timbre.

Attack and decay also contribute to the timbre of sound in some instruments. For example plucking a stringed instrument gives its sound a sudden attack which is characterized by a rapid rise to its peak amplitude. The decay is long and gradual by comparison. The ear is sensitive to attack and decay rates and uses them to identify the instrument producing the sound. In our research, we calculate the log attack time to capture this feature.

Monophonic sound means a sound having a single unaccompanied melodic line [60], which usually only has one instrument sound. **Polyphony** is music that simultaneously combines two or more independent musical lines (two melodies or a melody and a harmony), which results in multi-timbre sound with two or more instruments playing at the same time.

## *Single classifier on all features*

In k-nearest-neighbor prediction, the training data set is used to predict the value of a variable of interest for each member of a "target" data set. The structure of the data is such that there is a variable of interest (e.g., the instrument) and a number of conditional features. It is a so-called lazy learning model, by which

training is not necessary and learning is extremely fast. Its drawbacks include that  $k$  is an empirical value, which needs to be tuned among different classes of sounds.

Martin [18] employed the K-NN algorithm to a hierarchical classification system with 31 features extracted from cochleagrams. With a database of 1023 sounds they achieved 87% of successful classifications at the family level and 61% at the instrument level when no hierarchy was used. Using the hierarchical procedure increased the accuracy at the instrument level to 79% but it degraded the performance at the family level (79%). Without including the hierarchical procedure performance figures were lower than the ones they obtained with a Bayesian classifier. The fact that the best accuracy figures are around 80% and that Martin settled into similar figures, can be interpreted as an estimation of the limitations of the K-NN algorithm (provided that the feature selection has been optimized with genetic or other kind of techniques). Therefore, more powerful techniques should be explored.

Bayes Decision Rules and Naive Bayes classifiers are simple probabilistic classifiers, by which the probabilities for the classes and the conditional probabilities for a given feature and a given class are estimated based on their frequencies over the training data. They are based on probability models that incorporate strong independence assumptions, which often have no bearing in reality, hence are naive. The resultant rule is formed by counting the frequency of various data instance, and can be used then to classify each new instance. Brown [3] applied this technique to 18 Mel-Cepstral Coefficients by a K-means clustering algorithm and a set of Gaussian mixture models. Each model was used to estimate the probabilities that a coefficient belongs to a cluster. Then probabilities of all coefficients were multiplied together and were used to perform the likelihood ratio test. It then classified 27 short sounds of oboe and 31 short sounds of sax with an accuracy rate of 85% for oboe and 92% for sax.

Neural networks process information with a large number of highly interconnected processing neurons working in parallel to solve a specific problem. Neural networks learn by example. Cosi [6] developed a timbre classification system based on auditory processing and Kohonen self-organizing neural networks. Data was preprocessed by peripheral transformations to extract perception features, then fed to the network to build the map, and finally were compared in clusters with human subjects similarity judgments. In the system, nodes were used to represent clusters of the input spaces. The map was to generalize similarity criteria even to vectors not utilized during the training phase. All 12 instruments in the test could be quite well distinguished by the map.

Tree Classifiers Binary Tree is a data structure in which each node contains one parent and not more than 2 children. It has been pervasively used in classification and pattern recognition research. Binary Trees are constructed top-down with the most informative attributes as roots to minimize entropy. An adapted Binary Tree [14] was proposed with real-valued attributes for instrument classification regardless of pitch of the instrument in the sample.

Different classifiers for a small number of instruments have been used in music instrument estimation domain in the literature; yet it is a nontrivial problem to choose the one with optimal performance in terms of estimation rate for most western orchestral instruments. It is common to apply the different classifiers on the training data based on the same group of features extracted from raw audio files and get the winner with highest confidence for the unknown music sounds. The drawbacks include averaging the estimation efficiency by the tradeoffs among the features.

## **Multiple classifiers on different features**

Boosting systems [28] [29], based on multiple classifiers, achieve a better estimation model by training each given classifier on a different set of samples from training database, which keeps all the features or attributes. However music data usually could not take full advantage of such panel of learners because none of the given classifiers would get a majority weight, which is related to confidence, due to the homogeneous characteristics across all the data samples in training database. Thus the improvement can not be achieved by such combination of a number of classifiers.

Due to the existence of different characteristics for different features, the authors introduce a new method applicable to the music domain, which is to train different classifiers on different feature sets instead of different data samples. For instance, both MFCC and harmonic peaks are composed of serial real values, which are in form of numeric vectors and therefore work well with KNN instead of Decision tree. On the other hand, features such as zero crossing, spectrum centroid, roll-off, attack-time and so on, are acoustic features in form of single values, which could be combined to produce better rules after applied with decision tree or Bayes Decision Rules.

## Timbre relevant Features

The process of feature extraction is usually performed to extract structured data attributes from the temporal or spectral space of the signal. This will reduce the raw data into a smaller and simplified representation while preserving the important information for timbre estimation. Sets of acoustical features have been successfully developed for timbre estimation in monophonic sounds where mono instrument are playing.

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

Table.1 Feature

Group	Feature description
A	Spectrum Band Coefficients
B	MFCC
C	Harmonic Peaks
D	Spectrum Projection coefficients

**Spectrum Basis Functions** These functions, noted as  $\chi_k$ , are used to reduce the dimensionality by projecting the spectrum from high dimensional space to low dimensional space with compact salient statistical information.  $x_t$  is a vector of power spectrum coefficients in a frame t, which are transformed to log scale and then normalized. N, the total number of frequency bins, is 32 in 1/4 octave resolution.

Let  $V = [v_1 \ v_2 \ \dots \ v_k]$ , where  $V$  is computed from the equation below ( $USV$  is the function of standard singular decomposition, for detail, see Press et al. 1992).

$$\tilde{X} = USV^T$$

where,

$$\tilde{X} = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \vdots \\ \tilde{x}_M^T \end{bmatrix}$$

and for any  $i$ ,  $1 \leq i \leq M$ ,

$$\tilde{X}_i = \frac{\chi_i}{r} \quad \text{and}$$

$$r = \sqrt{\sum_{i=1}^N \chi_i^2} \quad \text{and}$$

$$\chi_i = 10 \log_{10}(x_i).$$

Since  $V$  is a matrix, statistical value retrieval has been performed for traditional classifiers.

**Spectrum Projection Functions** is a vector representing a reduced feature set by the projection against a reduced rank basis and it is computed by the formula:

$$y_t = \left[ r_t \quad \tilde{x}_t^T v_1 \quad \tilde{x}_t^T v_2 \quad \cdots \quad \tilde{x}_t^T v_k \right]$$

**Harmonic Peaks** is a sequence of local peaks of harmonics of each frame.

$$A(i, \text{harmonic}) = \max_{m \in [a, b]} (|X(m, i)|) = |X(M, i)|$$

$$f(i, \text{harmonic}) = M \times DF$$

$$a = \text{floor} \left( (\text{harmonic} - c) \frac{f0}{DF} \right), \quad b = \text{ceil} \left( (\text{harmonic} + c) \frac{f0}{DF} \right)$$

Where  $f0$  is the fundamental frequency in the  $i^{\text{th}}$  frame,  $\text{harmonic}$  is the order number of a harmonic peak,  $DF$  is the size of the frequency bin, where the total number of the frequency bin is  $NFFT$  ( $NFFT$  is the Next larger integer, which is a power of two. For example, the  $NFFT$  for 928 is 1024. ),  $c$  is the coefficient of the search range which is set to 0.10 in this paper.

**Mel frequency cepstral coefficients** describe the spectrum according to the human perception system in the mel scale [16]. They are computed by grouping the STFT (Short Time Fourier Transform) 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). We use the MFCC functions from the Julius software toolkit [1].

## Experiments

In order to validate the previous assumption, the authors built a database containing more than 4000 music instrument sounds which are taken from the McGill University Master Samples, and after segmenting those sounds into small slices (frames), we extracted the above features for each frame and saved them as the training and testing database.

Three experiments of classification based on the KNN and Decision Tree were conducted: 1) with all features; 2) with each feature group; 3) with the combination of different feature groups.

The feature retrieval system was implemented in C++. We used WEKA for all classifications. The training dataset of middle C includes 2762 records in our feature database. The frame-wise features are extracted from the following 26 instruments:

Electric Guitar, Bassoon, Oboe, B-flat clarinet, Marimba, C-Trumpet, E-flat Clarinet, Tenor Trombone, French horn, Flute, Viola, Violin, English horn, Vibraphone, Accordion, Electric Bass, Cello, Tenor saxophone, B-Flat Trumpet, Bass flute, Double bass, Alto flute, Piano, Bach trumpet, Tuba, and Bass Clarinet.

Due to the fact that sound features that represent various characteristics of timbre may have different degree of information loss during different classifier construction processes, we carried out three experiments to evaluate the features against the classifiers.

### ***Experiment I: Classification of all features***

In experiment I, we combined all the features (A to D) together as one single vector and applied KNN and Decision Tree (DT) classifiers to such vector database. J48 which is a pruned C4 algorithm was chosen as the decision tree classifier, confidence factor used for pruning was set as 0.25, and minimum number of instances per leaf as 2. As for KNN, we used Euclidean distance as the similarity function and assigned the K which is the number of neighbors as 3. All the features have been normalized by mean and standard deviation. 10-folder crossing validation was used for each classifier and the average confidence (accurately classified rate) was calculated, which is shown in Table2.

Table.2 Classification of all features

Classifier	Confidence (%)
KNN	98.22
DT	99.02

From the result, decision Tree shows a slightly higher confidence than KNN; however, there is no significant difference between KNN and DT.

### ***Experiment II: Classification of each feature***

In experiment II, the same process was performed except that classifiers are applied to each single feature database separately.

Table.3 Classification of each feature group

Feature Group	Classifier	Confidence (%)
A	KNN	99.23
	J48	94.69
B	IBK	98.19
	J48	93.57
C	IBK	86.60
	J48	91.29
D	IBK	47.45
	J48	31.81

The results in Table3 show that some features fit KNN better, such as band-coefficient, MFCC, projections while harmonic peaks has higher confidence under the decision tree classification.

### ***Experiment III: Classification of the combinations of different features***

In experiment III, we further combined every two features into a bigger feature vector and applied different classifiers respectively. The results are shown in the following figures, where y-axis indicates the confidence of classification, x-axis indicates the different feature or combinations of features.

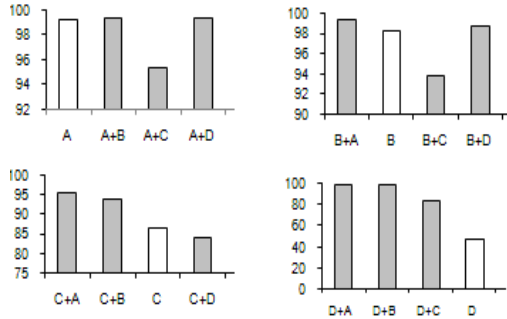


Fig.1 KNN classification in experiment III

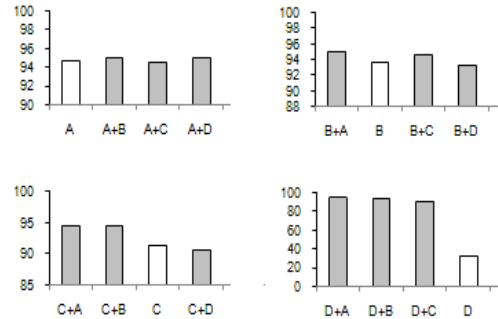


Fig.2 Decision tree classification in experiment III

Figure 2 shows that the confidence of classifier KNN tends to slightly go up as more features added. Yet when band-coefficient (which is feature A) is combined with harmonic-Peaks (which is feature C in the figure) the confidence significantly decreases. The same thing happened to the other features when they were combined with harmonic Peaks, which proved that KNN is less efficient for harmonic peaks than the other features. If the classification of KNN is constructed by the same dataset containing such features, the result tends to be deteriorated. Figure 3 shows that, for all the features with higher confidence in KNN, the accuracy does not change much in decision tree classification when they are combined with each other. And also when other groups are combined with Harmonic Peaks, there is no such significant decrease in confidence which observed in Table3.

We conclude that the KNN is more sensitive to the feature selection than decision tree in our music instrument classification. We also observed that harmonic peaks fit decision tree better than KNN in spite of its characteristic of multi-dimensional numeric vector which is similar to the other KNN-favored features. By adding more classifiers to the MIR system for estimating timbre with respective feature sets for the same audio objects, the system may improve its confidence in recognizing all the instruments in the database.

## MIR Framework based on multiple classifiers and features

Fig.11 shows the new strategy with a panel of classifiers applied on different feature sets of the same training data and the MIR system will benefit from the expertise of these classifiers in terms of accuracy and robustness.

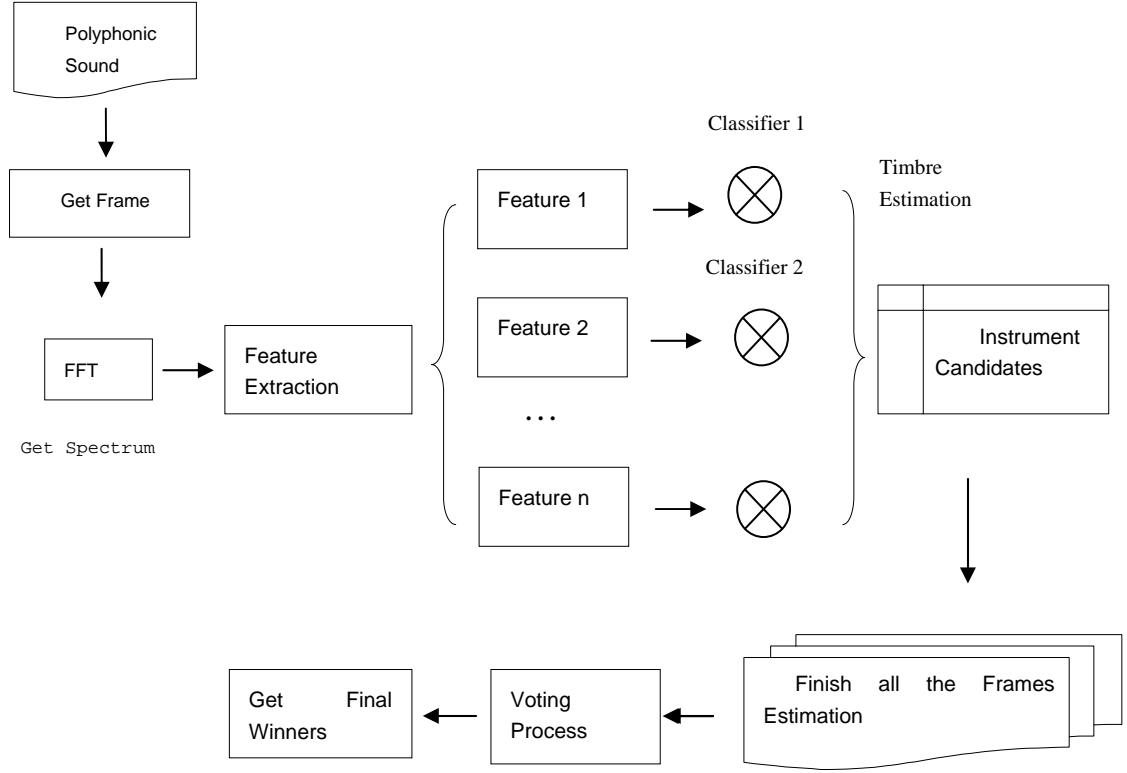


Fig.3 Timbre estimation with multiple classifiers and features

Let  $S = \{X, F, C, D\}$  be the multiple-classifier timbre estimation system, where the input analyzed audio sound is segmented into small frames  $X = \{x_1, \dots, x_t\}$ ,  $D = \{d_1, \dots, d_n\}$  are all the possible musical instrument class labels,  $F = \{f_1, \dots, f_m\}$  is the set of feature vectors which we extracted from the training database to build the classifiers  $C = \{c_1, \dots, c_m\}$ , and these features are also extracted from each analyzed frame to be classified by the classifiers respectively. Assume  $\lambda_1$  is the threshold for confidence which is the probability of the correct classification,  $\lambda_2$  is the threshold for support, the classification result of each classifier should satisfy these two thresholds

Thus, for each frame  $x_i$  where  $1 \leq i \leq t$ , we will get the instrument estimation  $d = C_j(f_j)$ , where  $d \in D, 1 \leq j \leq m, conf(d) \geq \lambda_1$  and  $sup(d) \geq \lambda_2$ . After evaluating all the frames, we get the overall

confidence for each instrument by summing up the confidence  $W(d_p) = \sum_{q=1}^t conf(d_p)_q$ , where  $1 \leq p \leq n$ , and

the final ranking and voting process is proceeded according to the weights  $W(d_p)$ . The top K musical instruments with highest overall confidence are selected as the final winners.

### ***Hierarchical structure of decision attributes***

According to how the sound is initially produced, the musical instruments are divided into different groups or families. The most commonly used system in the west today divides instruments into string instruments, wind instruments and percussion instruments. Erich von Hornbostel and Curt Sachs published an extensive new scheme

for classification. Their scheme is widely used today, and is most often known as the Hornbostel-Sachs system. The system 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. In this research, we do not discuss the membranophones family due to the lack of harmonic patterns in drums. Fig 4 shows the simplified Hornbostel/Sachs tree.

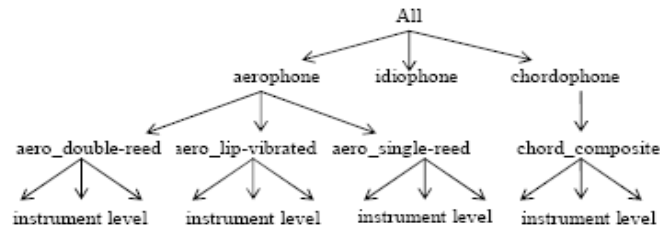


Fig.4 Homboch/sachs hierarchical tree

Fig 5 shows us another tree structure of instrument families which is grouped by the way the musical instruments are played. We will later use these two hierarchical trees as the samples to introduce the cascade classification system and give the testing results.

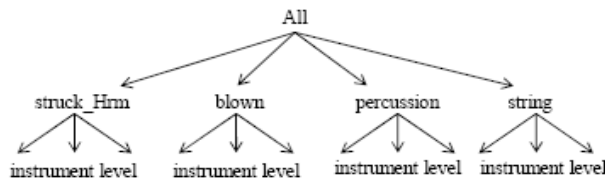


Fig.5 Play method hierarchical tree

According to the experience of human's recognition of musical instruments, it is usually easier for one person to tell the difference among the instruments when those instruments belong to the different families than to distinguish those which belong to the same family. For instance, violin and piano each belong to aerophone and chordophone in the Homboch/Sachs structure. And in play-method structure, they each belong to blown family and struck family. So it makes their tone color or sound quality sound quite different from each other which lead to easier identification of the two instruments in polyphonic sound. However, when it comes to distinguishing the violin from the viola, people need to pay more attention to discern each of them since both instruments fall into the same category of string instruments in play-method structure and chordophone family in Homboch/Sachs structure which indicates that they produce similar timbre. So if we build the classifiers on each level of these hierarchical decision structures, the classifier of the higher level is to be applied in order to get the estimation of the instrument family, then the classifier of the lower level is applied to analyze the musical sound in order to further narrow down the range of possible instruments. The cascade classification process is performed from the root toward the bottom along hierarchical tree, until it reaches the bottom level which gives the specific instrument name estimation. The classifiers of the lower level are built on the subset of the training data which corresponds to the particular instrument family, which means the classifiers are specifically trained for the purpose of identifying a smaller number of instruments with a small family range and thus give them expertise to better fits the estimation task of instruments which fall in this particular family.

## Cascade classifier of Hierarchical Decision Systems

To verify the assumption of the advantage of the cascade classification system, the authors built a multi-hierarchical decision system  $S$  with all the low-level MPEG7 descriptors as well as other popular descriptors for describing music sound objects. The decision attributes in  $S$  are hierarchical and they include Hornbostel-Sachs classification and classification of instruments with respect to playing method.

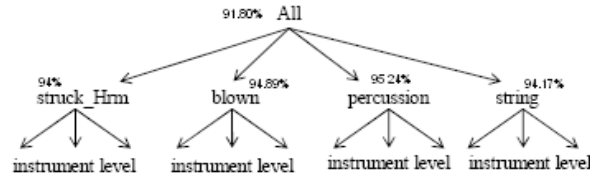


Fig.6 Cascade classifier for classification of instruments with respect to playing method and their confidence

The information richness hidden in the descriptors has strong implications on the confidence of classifiers built from  $S$ . Hierarchical decision attributes allow us to have the indexing done on different granularity levels of classes of music instruments. We can identify not only the instruments playing in a given music piece but also classes of instruments if the instrument level identification fails.

In this section we show that cascade classifiers outperform standard classifiers. The first step in the process of recognizing a dominating musical instrument in a musical piece is the identification of its pitch. If the pitch is found, then a pitch-dedicated classifier is used to identify this instrument.

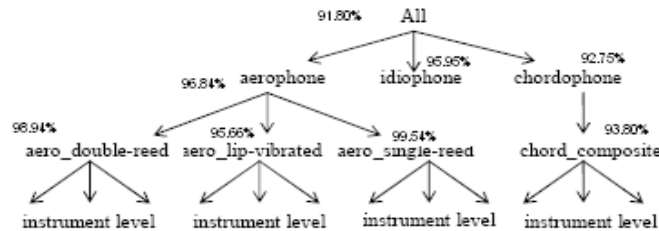


Fig.7 Cascade classifier for Hornbostel-Sachs classification of instruments and their confidence

The testing was done for music instrument sounds of pitch 3B. The results are shown in Figure 6 and Figure 7. The confidence of a standard classifier  $\text{class}(S, d, 3)$  for Hornbostel-Sachs classification of instruments is 91.50%. However, we can get much better results by following the cascade approach. For instance, if we use the classifier  $\text{class}(S, d, 2)$  followed by the classifier  $\text{class}(S, d[1, 1], 3)$ , then its precision in recognizing musical instruments in aero double reed class is equal to  $96.02\% * 98.94\% = 95.00\%$ . Also, its precision in recognizing instruments in aero single reed class is equal to  $96.02\% * 99.54\% = 95.57\%$ . It must be noted that this improvement in confidence is obtained without increasing the number of attributes in the subsystems of  $S$  used to build the cascade classifier replacing  $S$ . Clearly, if we increase the number of attributes in these subsystems then the resulting classifiers forming the cascade classifier may easily have higher confidence and the same the confidence of the cascade classifier will be increased.

Looking again at Figures 6 and 7, when we compare different classifiers which are built on the same training dataset but on a different level of decision value based on our hierarchical trees, we found that generic classifiers usually have higher recognition accuracy than the peculiar one.

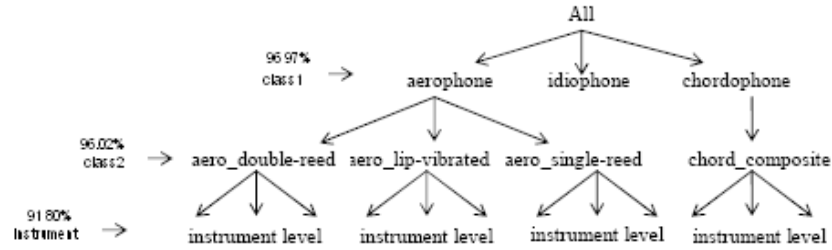


Fig.8 The accuracy of classifiers built on different level of decision attributes (pitch 3B)

By this strategy, we are getting higher accuracy for single instrument estimation than the regular method. As we can see, the accuracy has reached the point which would minimize the effects of mismatching multiple instrument patterns due to the similarity among them.

### Feature and classifier selection at each level of cascade system

In order to get the highest accuracy for the final estimation at the bottom level of the hierarchical tree, the cascade system must be able to pick a feature and a classifier from the available features pool and classifiers pool in such a way that the system achieves the best estimation at each level of cascade classification. To get such information, we need to deduce the knowledge from the current training database by combining each feature from the feature pool (A, B, C, D) with each classifier from the classifier pool (NaiveBayes, KNN, Decision Tree), and running the classification experiments in Weka on the subset which corresponds to each node in the hierarchical tree used by cascade classification system.

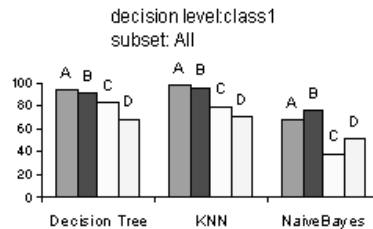


Fig.9 Classification on top level with different classifiers

Figure 9 shows that on the top level, KNN and feature A got the highest estimation when the decision level was on class1, which means at the beginning the system should use band coefficients as the feature to run the KNN classification algorithm to find which family the target object belongs to. In order to go further to the second level of the tree, the system has to decide the pair selection of feature-classifier based on the following knowledge derived from classification results running on the different subsets of training data at the second level.

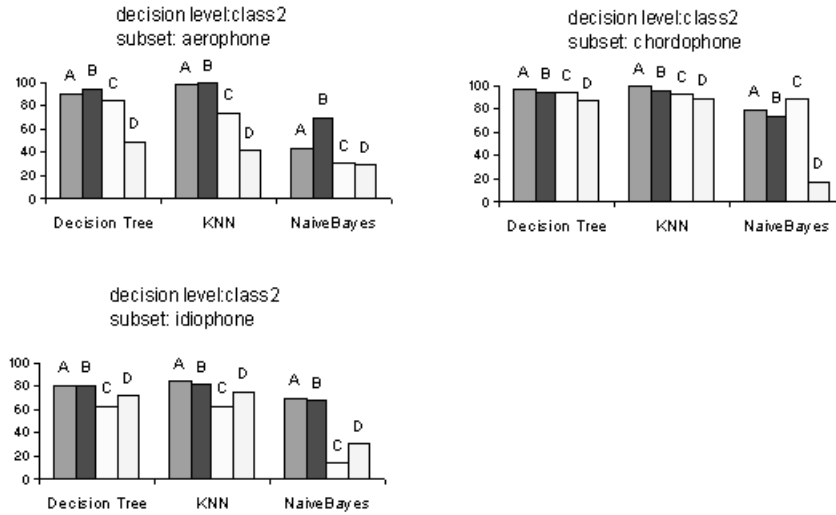


Fig.10 Classification on second level with different classifiers

From Figure 10, we can see that KNN classifier and feature A(band coefficients) is still the best choice for the subsets of Chordophone and Idiophone. Yet feature B(MFCC) outperformed any other features in the group of Aerophone. Table.4 shows such conclusion more clearly.

Table.4 Feature and classifier selection table for Level1

Node	feature	Classifier
chordophone	Band Coefficients	KNN
aerophone	MFCC	KNN
idiophone	Band Coefficients	KNN

Again, we continue to perform the classification on the different subsets of training data at third level subsets of Hornbostel-Sachs hierarchical tree and get the classification confidence results as the Figure 11 shows.

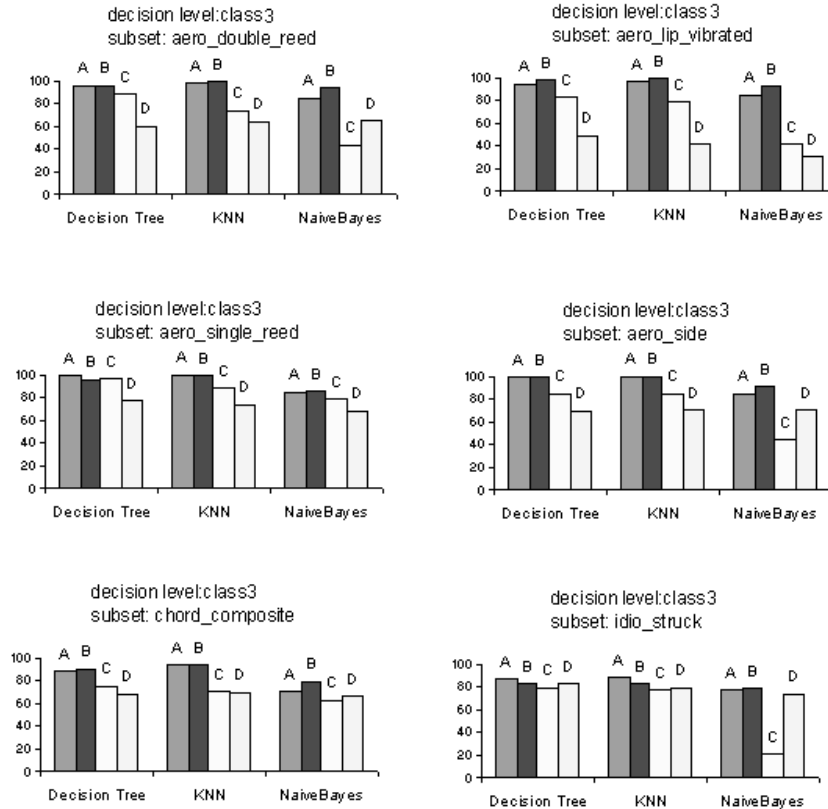


Fig.11 Classification on third level with different classifiers

The instrument name is eventually estimated by the classifiers at the third level. We also observed some interesting results of the classifier and feature selection. The subset of Aero\_single\_reed does not inherit the characteristic from the parent node (Aerophone) as the other Aerophone subsets (aero\_double-reed, aero\_lip-vibrated, aero\_side) do. Decision tree along with Feature A(Band Coefficients) has the highest confidence instead of Feature B(MFCC) with KNN. Table 5 shows the details of the best choice of feature selection and classifier selection.

Table.5 feature and classifier selection table for Level2

Node	feature	Classifier
chrd_composite	Band Coefficients	KNN
aero_double-reed	MFCC	KNN
aero_lip-vibrated	MFCC	KNN
aero_side	MFCC	KNN
aero_single-reed	Band Coefficients	Decision Tree
idio_struck	Band Coefficients	KNN

From these results, we concluded that the classification confidence could be improved in cascade classification system by choosing the appropriate feature and classifier at each level of hierarchical tree.

## MIR Framework based on cascade classification system

Fig.12 shows another framework based on feature selection and classifier selection in the cascade hierarchical classification system. The system will perform timbre estimation for polyphonic sound with high accuracy while still preserving the applicable analyzing speed by choosing the best feature and classifier for the classification process at each level based on the previous knowledge derived from the training database.

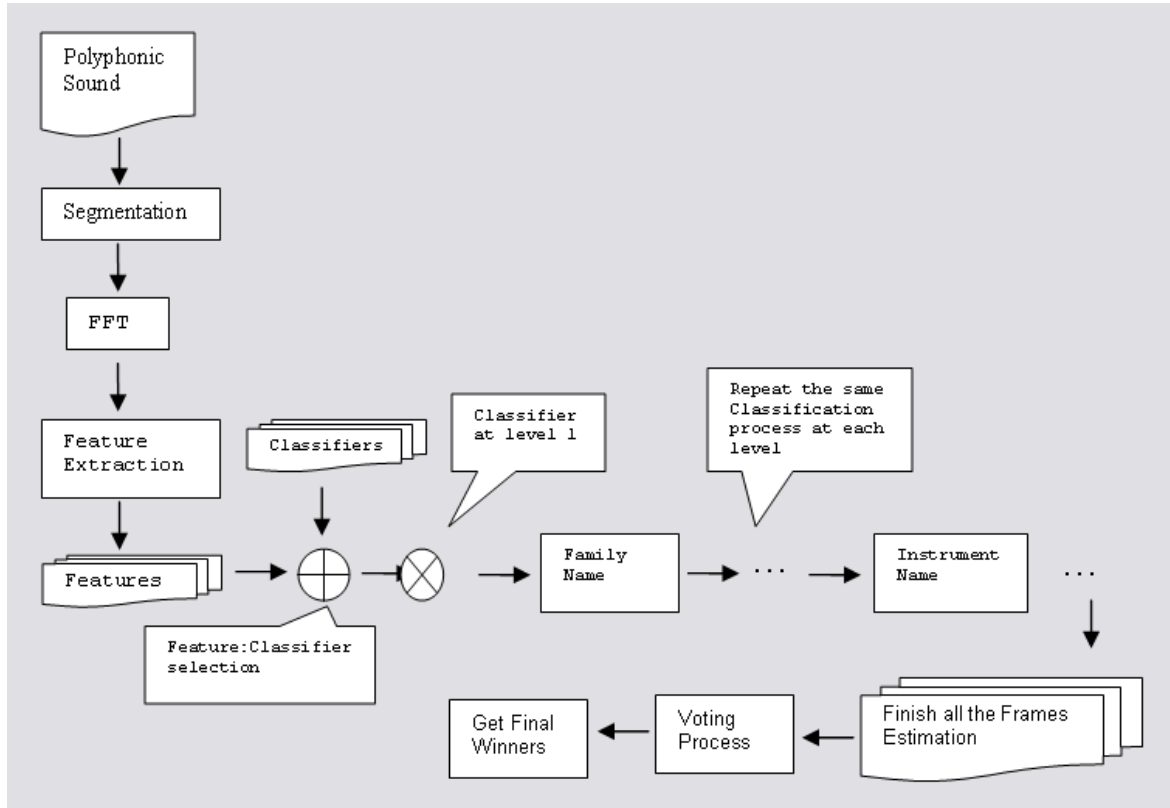


Fig.12 Timbre estimation with classifier and feature selection

Let  $S = \{X, F, C, D, L\}$  be the multiple-classifier timbre estimation system, where the input analyzed audio sound is segmented into small frames  $X = \{x_1, \dots, x_t\}$ ,  $D = \{d_1, \dots, d_n\}$  are all the possible musical instrument class labels,  $F = \{f_1, \dots, f_m\}$  are the feature vectors which are extracted from training database to build the classifiers  $C = \{c_1, \dots, c_w\}$ , and these features are also extracted from each analyzed frame to be classified by the classifiers respectively.  $L = \{l_1, \dots, l_v\}$  Assume  $\lambda_1$  is the threshold for confidence which is the probability of the correct classification,  $\lambda_2$  is the threshold for support, the classification result of each classifier should satisfy these two thresholds.

Thus, for each frame  $x_i$  where  $1 \leq i \leq t$ , at each level  $\alpha$  of cascade system, we will have the pair of  $(C_z, f_y)$ , where  $1 \leq z \leq m$ ,  $1 \leq y \leq w$ , and get the estimation confidence  $conf(x_i, \alpha) = C_z(f_y)$ , where  $d \in D, 1 \leq j \leq m, conf(x_i, \alpha) \geq \lambda_1$  and  $sup(x_i, \alpha) \geq \lambda_2$ . After evaluating all the levels, we get the final instrument name estimation  $d_p$ , where  $d \in D$ , and the final confidence for the instrument by multiplying

the confidence of each classification level for the frame  $x_i$ ,  $conf(x_i, d_p) = \prod_{\alpha=1}^v conf(x_i, \alpha)$ . After all the

frames are classified, the overall weights for each estimated instrument is calculated by  $W(d_p) = \sum_{q=1}^t conf(d_p)_q$

where  $1 \leq p \leq n$ . Then the ranking and voting process is preceded according to the weights  $W(d_p)$ . The top  $K$  musical instruments with highest overall confidence are selected as the final winners, where  $K$  is the parameter assigned by the user.

## Conclusion and future work

We conclude that the KNN algorithm is more sensitive to feature selection than decision tree in our music instrument classification. We also observed that the harmonic peaks feature fits decision tree better than KNN in spite of the fact that it is a multi-dimensional numeric vector which is similar to the other KNN-favored features. By adding more classifiers to the MIR system to estimate timbre with respective feature sets for the same audio objects, the system could have a higher confidence for all the instruments in the database. Future work includes investigating more classifiers such as Support Vector Machines, Naive Bayes, and Neural Networks to get better knowledge of their expertise in different feature sets. Also the testing on the MIR system with the proposed new strategy with multiple classifiers on different features needs to be performed to further prove the improvement of the robustness and recognition rate of timbre estimation for polyphonic music sound.

Because two previous hierarchical structures try to group the instruments according to the semantic similarity proposed by human experts, quite often the instruments are assigned to the same group even their sounds are quite different. On the other hand, two instruments can be assigned by the hierarchical structure to different groups even though they have similar sound quality which clearly may confuse the timbre-estimation system. For instance, trombone belongs to the aerophone family; however, system often classifies it as chordophone, such as violin. This is because of the inherent falsehood and ambiguousness that exists in those instrument categories. In order to make the hierarchy structure fit the feature-based classification system, we will build a new family tree for musical instruments by clustering them in the way the instruments in the same family have the same sound quality from the perspective of the machine. We will apply clustering algorithms such as EM and k-means to regroup the instruments by the similarity of features which are also used for timbre estimation.

## Acknowledgment

This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-0414815. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## Reference

- [1] A survey of music information retrieval systems, <http://mirsystems.info/>
- [2] Akinobu, LEE et al. Julius software toolkit. (<http://julius.sourceforge.jp/en/>)
- [3] Brown, J. C. (1999). Musical instrument identification using pattern recognition with cepstral coefficients as features, Journal of Acoustical society of America, 105(3), 1933–1941.

- [4] Brown, J. C. (2001). Houix, O., McAdams, S., Feature dependence in the automatic identification of musical wind instruments, in J. Acoust. Soc. of America, 109, 2001, 1064-1072.
- [5] Bregman, A.S. (1990). Auditory scene analysis, the perceptual organization of sound, MIT Press
- [6] Cosi, P. (1998). Auditory Modeling and Neural Networks, in A Course on Speech Processing, Recognition, and Artificial Neural Networks, Springer Verlag, Lecture Notes in Computer Science, in fase di stampa.
- [7] Cutting D., Kupiec, J., Jan Pedersen, and Penelope Sibun, (1992). A Practical Part-of-Speech Tagger, in the Third Conference on Applied Natural Language Processing, pp. 133-140.
- [8] Czyzewski, A. (1998). Soft processing of audio signals, in Polkowski, L. and Skowron, A. (eds.) Rough Sets in Knowledge Discovery Heidelberg: Physica Verlag, pp. 147-165.
- [9] Kaminskyj, I. (2000). Multi-feature Musical Instrument Classifier, MikroPolyphonie 6, 2000 (online journal at <http://farben.latrobe.edu.au/>)
- [10] Kostek, B. (1998). Soft computing-based recognition of musical sounds, in Polkowski, L. and Skowron, A. (eds.) Rough Sets in Knowledge Discovery Heidelberg: Physica-Verlag.
- [11] Kupiec, J. (1992). Robust Part-of-Speech Tagging Using a Hidden Markov Model. In the Computer Speech and Language 6, pp. 225-242.
- [12] Kostek, B. and Czyzewski, A. (2001). Representing Musical Instrument Sounds for Their Automatic Classification, in J. Audio Eng. Soc., Vol. 49, No. 9, 2001, 768-785
- [13] Herrera, P., Amatriain, X., Batlle, E., Serra X. (2000). Towards instrument segmentation for music content description: a critical review of instrument classification techniques, in the international Symposium on Music Information Retrieval (ISMIR 2000), Plymouth, MA, 2000.
- [14] Jensen, K., Arnspang, J. (1999) Binary decision tree classification of musical sounds, the 1999 International Computer Music Conference, Beijing, China, Oct.
- [15] Lindsay, A. T., and Herre, J. (2001) MPEG-7 and MPEG-7 Audio—An Overview, J. Audio Eng. Soc., vol.49, July/Aug, pp. 589–594.
- [16] Logan, B. Mel (2000). Frequency Cepstral Coefficients for Music Modeling, in Proc. 1st Ann. Int. Symposium On Music Information Retrieval (ISMIR).
- [17] Martin, K. D. (1999). Sound-Source Recognition: A Theory and Computational Model, Ph.D. Thesis, MIT, Cambridge, MA.
- [18] 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.
- [19] Paulus, J., Virtanen, T. (2005). Drum transcription with non-negative spectrogram factorization, Proceedings of 13. European Signal Processing Conference, EUSIPCO, Antalya, Turkey, 4-8 September 2005
- [20] Polkowski, L. and Skowron, A. (1998). Rough Sets in Knowledge Discovery Heidelberg: Physica-Verlag.
- [21] Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). Numerical Recipes in C (2<sup>nd</sup> edition). Cambridge.
- [22] Ras, Z. and Wieczorkowska, A. (2001). Indexing audio databases with musical information, A., in Proceedings of SCI'01, Volume 10, Orlando, Florida, July 22-25, 2001, 279-285.

- [23] Scheirer, E. and Slaney, M. (1997). Construction and Evaluation of a Robust Multi-feature Speech/Music Discriminator, in Proc. IEEE int. Conf. on Acoustics, Speech and Signal Processing (ICASSP).
- [24] Tzanetakis, G. and Cook, P. (2002) "Musical Genre Classification of Audio Signals," IEEE Trans. Speech and Audio Processing, July, vol. 10, pp. 293-302.
- [25] 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.
- [26] Wieczorkowska, A. and Ras, Z. (2001). Audio content description in sound databases, in Web Intelligence: Research and Development, WI'01, Maebashi City, Japan, LNCS/LNAI 2198, Springer-Verlag, 175-183
- [27] Wold, E., Blum, T., Keislar, D., and Wheaton, J., (1996). Content-Based Classification, Search and Retrieval of Audio, IEEE Multimedia, Fall, pp. 27-36.
- [28] Yoav Freund. Boosting a weak learning algorithm by majority. Proceedings of the Third Annual Workshop on Computational Learning Theory. 1990
- [29] Yoav Freund and Robert E. Schapire A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119--139, 1997
- [30] Young, S.J., Russell, N.H., and Thornton, J.H. (1989). Token passing: a simple conceptual model for connected speech recognition systems. Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Department, Cambridge, UK, July.
- [31] Zhang, X., Marasek, K., and Ras, Z.W. (2007). Maximum Likelihood Study for Sound Pattern Separation and Recognition, in proceedings of the IEEE CS International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), April 26-28, in Seoul, Korea, 807-812.