Instrument Classification in Polyphonic Music
Based on Timbre Analysis

Tong Zhang
Hewlett-Packard Laboratories
1501 Page Mill Road, Palo Alto, CA 94304
Email: tong_zhang@hp.com

ABSTRACT

While most previous work on musical instrument recognition is focused on the classification of single notes in monophonic music, a scheme is proposed in this paper for the distinction of instruments in continuous music pieces which may contain one or more kinds of instruments. Highlights of the system include music segmentation into notes, harmonic partial estimation in polyphonic sound, note feature calculation and normalization, note classification using a set of neural networks, and music piece categorization with fuzzy logic principles. Example outputs of the system are "the music piece is 100% guitar (with 90% likelihood)" and "the music piece is 60% violin and 40% piano, thus a violin/piano duet". The system has been tested with twelve kinds of musical instruments, and very promising experimental results have been obtained. An accuracy of about 80% is achieved, and the number can be raised to 90% if misindexings within the same instrument family are tolerated (e.g. cello, viola and violin). A demonstration system for musical instrument classification and music timbre retrieval is also presented.

Keywords: musical instrument classification, music timbre analysis, polyphonic music, note onset detection, music timbre features, harmonic partials, music retrieval, music database management.

1. INTRODUCTION

Nowadays there are more and more music in digital formats available, and they are becoming more and more popular in people's life. Then comes the problem of how to efficiently arrange musical assets so as to make them easy for browsing and retrieving. Since the volume of digital music materials is growing rapidly, manual indexing and retrieval are just impossible. The proposed system is part of our current work on automatic categorization and retrieval of musical assets based on instrument, rhythm and melody of the music. There are many applications of this work such as online music shopping, personal music library organization, searching for preferred music channels on the web, and retrieving video segments based on music content.

In this paper, we focus on identifying music pieces according to the instruments involved in the music. There is existing work, although quite limited, on instrument classification or music timbre recognition. The two problems are similar because the distinction of different instruments is largely determined by the different timbres they produce. Nevertheless, most previous researches are concerned with the classification of single notes in monophonic music, that is, there is only one instrument and one note played at one time. Furthermore, synthetic music (e.g. MIDI music) is usually used for experiments in such researches which makes it far away from real application scenarios. Techniques proposed in existing work mainly fall into two parts: feature extraction methods and classification tools. For the first part, music features which have been used for instrument distinction or timbre analysis include the note features [1],[2], cepstral coefficients [3], wavelet coefficients [4] and auditory modeling parameters [5]. For the second part, mostly used classification tools are the different variations of artificial neural networks [1],[4]-[8]. There are other pattern recognition methods applied such as rough sets [2] and Gaussian mixture models [3]. While in most papers only quite vague or simple results were presented, the outputs from [5] look more promising: by training Kohonen's self-organizing map (SOM), 12 kinds of instruments correspond to different regions in the map, and similar instruments (e.g. instruments in the same family) have regions close to each other. However, the auditory modeling method used in [5] is too complicated to be applied in real-time applications.

Compared with existing approaches, one major distinction of the proposed work is that we develop a system for real music including music downloaded from the web, recorded from TV or radio, and music tracks on compact discs, rather than synthetic music such as MIDI. The approach works with continuous music pieces instead of single notes. Also, we develop methods which apply for not only monophonic music but also polyphonic music (i.e. more than one instruments or notes played at one time), because it is very common that there are more than one kinds of instruments involved in a music piece. Moreover, the algorithms are fast enough for real-time applications. In general, the proposed approach is much more suitable for application scenarios in real life.

In the proposed system, we have considered twelve kinds of instruments for training and testing purposes, including cello, viola, violin and guitar from the string family; flute, horn and trumpet from the wind family; piano and organ from the keyboard family; and three kinds of oriental musical instruments used for comparison: erhu, zheng and sarod. The preliminary results are very promising: through experiments with a database containing 287 music pieces, an accuracy rate of about 80% is obtained. This number can be raised to 90% if misindexings within the same instrument family can be tolerated.

The rest of this paper is organized as follows. An overview of the proposed system is given in Section 2. Then, modules of the system including music piece segmentation, harmonic partial estimation, note feature computation and normalization, note classification and music piece classification are described in Sections 3 to 7, respectively. Experimental results and demonstrations are presented in Section 8, and finally, conclusions and plans for future work are given in Section 9.

2. SYSTEM OVERVIEW

The proposed system comprises six modules, as illustrated in Figure 1. A music piece is first segmented into notes by detecting note onsets. This is done based on analyzing the temporal energy envelope and its 1st order difference of the music piece. Then, for each note, the harmonic partials are estimated using an algorithm which applies for both clean and noisy, monophonic and polyphonic music. Next, features are computed for each note separately, including temporal features, spectral features and partial features. These audio features are recognized as being important in representing the timbre of one note. Some of the note features also need to be normalized so that they are independent of the volume, the length and the pitch of one note. A feature vector is then formed for each note which is to be sent to the classifier.

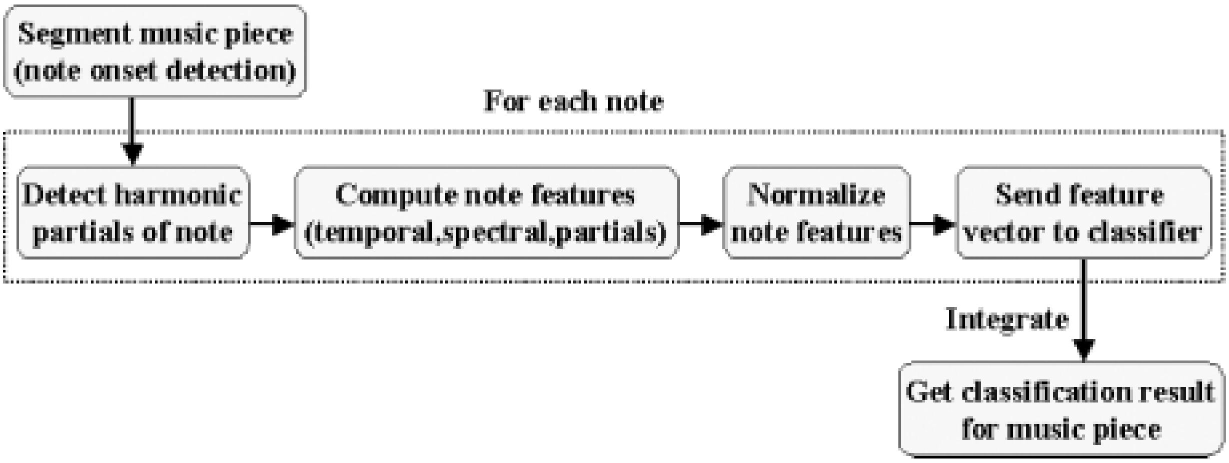
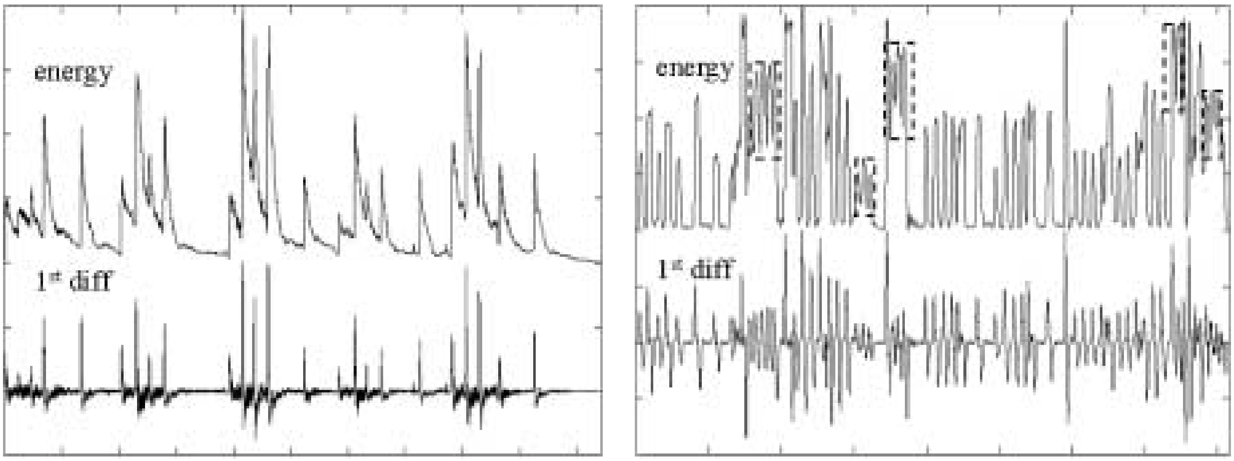

Segment music piece (note onset detection)
For each note
Detect harmonic partials of note
Compute note features (temporal,spectral,partials)
Normalize note features
Send feature vector to classifier
Integrate
Get classification result for music piece

Figure 1. Flowchart of the proposed system for musical instrument classification.

A set of classification tools are used to classify one note to one kind of instrument. The Kohonen's self-organizing map, which is a type of unsupervised artificial neural network, is employed to find the optimal structure of the feature vector. The multi-layer perceptron, a supervised neural network, is used to estimate the probability of one note's belonging to be a certain instrument. Finally, the classification results of all notes within a music piece are integrated to categorize the music piece to one or more kinds of musical instruments. In addition, the Gaussian mixture models (GMM) are used to add similarity search capability to a music database. That is, a GMM is built for each music piece in a database based on note feature vectors within the piece. Then, for a given music piece, those music samples which have similar timbre features to this piece can be retrieved by matching its note feature vectors to the existing GMMs in the database.

3. MUSIC PIECE SEGMENTATION

To segment a music piece into notes, the onset of each note (i.e. the starting point of a note) is detected. Plotted in Figure 2(a) and (b) are the temporal energy envelope (upper curve in each figure) and its 1st order difference (lower curve in each figure) of a piano piece and a flute piece, respectively. As can be observed, note onsets are indicated by sharp drop and/or rise of energy value in the temporal envelope of a music piece.



(a) (b)

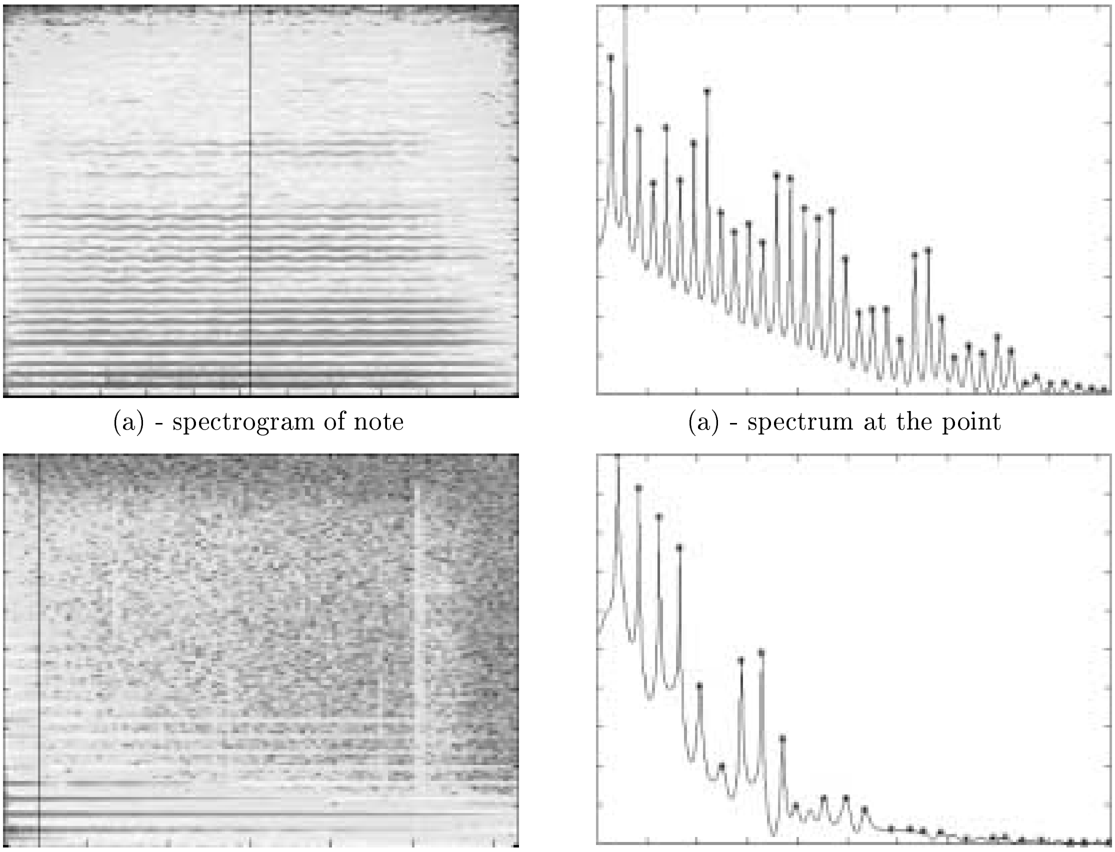
Figure 2. Temporal energy envelope and its 1st order difference of music pieces: (a)piano and (b)flute.

We denote the temporal envelope as E, and its 1st order difference as D. Potential note onsets are first obtained by applying a twin-threshold scheme to D. Two thresholds, T_h and T_l, are adaptively determined based on the mean of E and the standard deviation of D according to an empirical formula. T_h is higher than T_l by a fixed ratio. Then, those locations in D are marked as potential note onsets which either have a positive peak higher than T_h, or have a positive peak higher than T_l with a negative peak lower than $-T_h$ right before it.

Next, the exact locations for note onsets are found in E by searching for the lowest point within the attacking phase of a note. Meanwhile, false alarms are removed which are surplus potential onsets within one note, especially those caused by the vibration of instrument. Several notes having vibration are marked with dash-line blocks in Figure 2(b). After that, the ending point of a note is searched in E starting from the onset point. The range of each note is then determined. It should be noted that no matter how the thresholds are selected, it is never possible to find all note onsets with the temporal envelope alone. Instead, only those notes strong enough are detected and weak notes are ignored. This is actually good for the following modules and the overall classification accuracy, because partial detection and partial parameter calculation may be unreliable for weak notes.

4. HARMONIC PARTIAL ESTIMATION

The locations and amplitudes of harmonic partials in a note are important features indicating timbre of the note. Since in a continuous music piece, especially polyphonic music, the starting point and ending point of a note may be overlapped with part of the previous note or the next note, we choose to estimate harmonic partials only at one strong point for each note in this work. That is, the point with high energy value and at which most partials are present is selected. However, while for notes having sustaining phase such as in trumpet and violin, the strongest point is normally in the middle of the sustaining phase; for notes without sustaining phase like in piano and guitar, the strongest point is at the attacking phase. Thus, a scheme accommodating to both kinds of cases is designed as follows: if a note has length smaller than 300ms, choose the point right at the middle of the note; otherwise, denote the note onset as A, the point at 150ms as B and the middle point of the note as C, search for the point D between A and C which has the highest energy value, then choose the maximum of B and D. Two examples of finding the right point to estimate partials are shown in Figure 3. For both examples, the figure on the left is the spectrogram of the note, where the horizontal axis denotes time, the vertical axis denotes frequency, and the pixel luminance represents the energy level at that point. We can see the horizontal lines in the spectrogram which indicate harmonic partials in the note. The vertical line in the spectrogram stands for the point which was chosen to estimate partials. On the right, is the spectrum calculated at the selected point, and the stars indicate peaks in the spectrum which were detected by the peak-picking algorithm.



(a) - spectrogram of note

(a) - spectrum at the point

(b) - spectrogram of note

(b) - spectrum at the point

Figure 3. Find the right point in a note to estimate partials: (a)cello and (b)guitar.

The spectrum is generated with autoregressive (AR) model coefficients estimated from the autocorrelation of the audio signal. This AR model generated spectrum is a smoothed version of the frequency representation. Moreover, as the AR model is an all-pole expression, peaks are prominent in the spectrum. Overall, it is much easier to detect harmonic peaks in AR model generated spectrum than in the directly computed one (using only one FFT). The energy of the estimated spectrum is normalized so that note features are independent of the loudness or volume of the note. Prominent peaks are detected in the spectrum and their corresponding frequencies form a set P.

After that, a set of candidate values for the fundamental frequency (FuF) is generated as follows: for any peak P_k in P which is between 50Hz and 3000Hz, put P_k, $P_k/2$, $P_k/3$ and $P_k/4$ in a set F; rearrange F so that duplicate values are removed, and values outside the range of 50-2000Hz are removed. Then, for each candidate value F_k in F, search for its partials in P as below: $C_{k1} \sim F_k, C_{k2} \sim C_{k1} + F_k, \ldots, C_{ki+1} \sim C_{ki} + F_k, \ldots$, where C_{ki} is the ith partial to be searched. If C_{ki} is not found, then $C_{ki+1} \sim C_{ki-1} + 2F_k$, and so on. This procedure can avoid the accumulation effect from inaccuracy in the candidate FuF value. Next, a score $S(F_k)$ is computed based on the number and parameters of obtained partials of F_k, including amplitude, width and sharpness of partials, and using an empirical formula. After scores of all the members in F are calculated, choose the candidate F_m which has the largest score $S(F_m)$. If $S(F_m)$ is greater than a predefined threshold, then F_m and its corresponding partials are the estimated fundamental frequency and harmonic partials of the note; otherwise, the fundamental frequency value is set to zero, and there are no partials estimated for this note.

This method of detecting harmonic partials works not only with clean music, but also with music having noisy background; not only with monophonic music, but also with polyphonic music in which two or more instruments are played at the same time. Polyphonic music is quite common in music performances such as piano and violin duet, trumpet and organ duet, etc. In the case of polyphonic music, the note with the strongest partials, as indicated by the highest score computed, will be detected. In a music piece, different instruments may be dominant at different periods, and partials of the strongest instrument are always detected at any time.

5. NOTE FEATURE COMPUTATION AND NORMALIZATION

The timbre of a note is determined by audio features from various aspects, including the following [9]:

- temporal envelope of the note
- spectral power distribution of the note
- locations and amplitudes of the partials
- degree of inharmonicity of the partials
- presence of high frequencies at the moment of the attack
- whether the partials grow or decrease at the same time

Since it is difficult to obtain features of the attacking moment and the synchrony of partials in a continuous music piece, we ignore the last two aspects of features in this work, and only consider the first four types of features. The audio features are categorized into three groups as described below.

5.1. Temporal features

The temporal envelope of the note is computed, which can be divided into three phases: attacking, sustaining and releasing. Shown in Figure 4 are some examples of the temporal envelope of note. We can see that in the piano note, the energy rises quickly in the attacking phase, then releases slowly, and there is almost no sustaining phase. In the trumpet note, the energy also rises quickly, but it has a remarkable sustaining phase, and also releases faster than the piano note. In both the violin and organ notes, there are ripples in the sustaining phase which are caused by the vibration in the instrument. However, compared to that of the organ note, the energy of the violin note is slower in both the rising and the releasing speed.

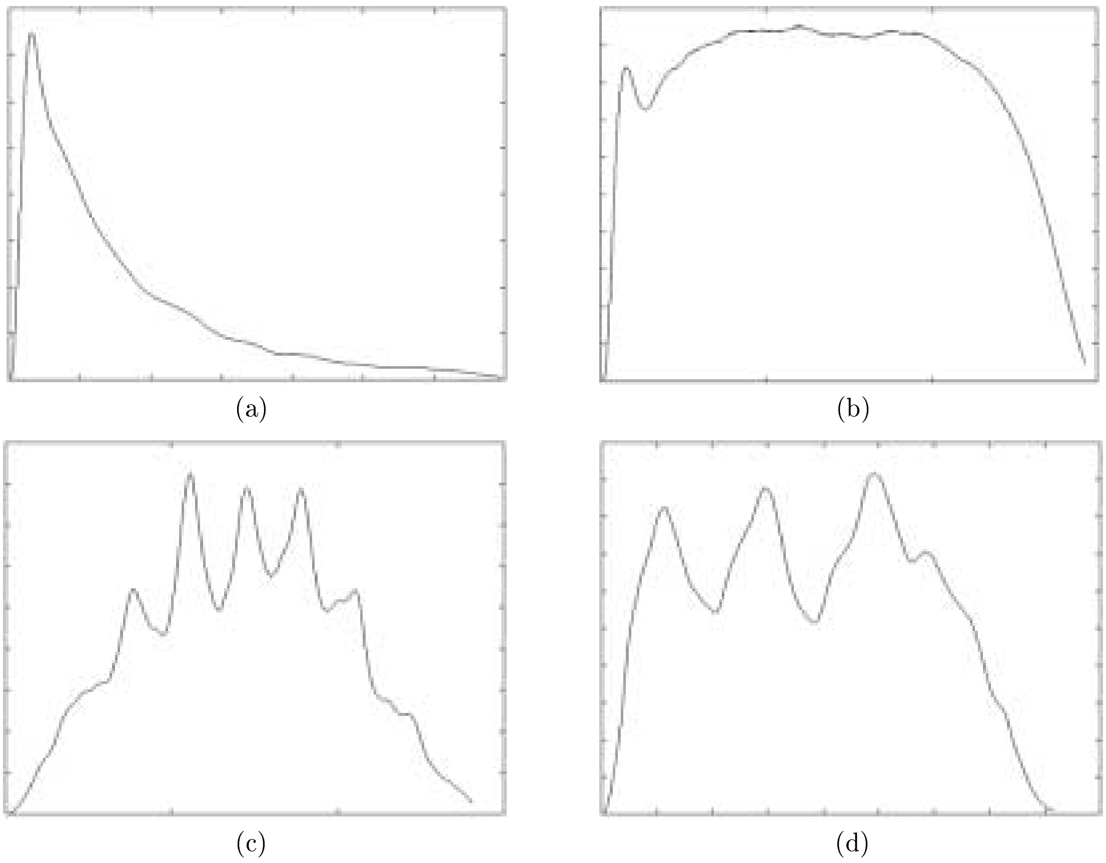
Four features of the temporal envelope are calculated as follows:

- Rising speed (S_a): the average slope in the attacking phase.
- Degree of sustaining (D_s): length of the sustaining phase.
- Degree of vibration (D_v): sum of amplitudes of prominent ripples in the sustaining phase.
- Releasing speed (S_r): the average slope in the releasing phase.

5.2. Spectral features

While some instruments generate sounds which have energy concentrated in the lower frequency bands, there are other instruments which produce sounds with energy almost evenly distributed among lower, mid, and higher frequency bands. Several examples are shown in Figure 5. Plotted in (a) and (b) are the spectra of a flute note and a trumpet note, respectively. While the two notes have the same loudness and pitch, the trumpet note has more energy distributed in the mid to high frequency bands than the flute note, which contributes to the distinction in the timbre of the two notes. Similar situation happens to the guitar note and the zheng note whose spectra are shown in (c) and (d), respectively. Partials in the high frequency band make the zheng note sound brighter than the guitar note.

In the spectrum of a note which is computed as described in the last section, we divide the frequency axis equally into three sub-bands. Suppose the total energy of all the partials in the spectrum is E_{sum}, and the total energy of partials in each sub-band is $E_k, k = 1 \sim 3$. Then, the ratios between each $E_k, k = 1 \sim 3$, and E_{sum} are calculated. These ratios represent the spectral energy distribution of the sound among sub-bands, and are denoted as $E_r = \{E_{r1}, E_{r2}, E_{r3}\}$.



(a) (b) (c) (d)

Figure 4. Temporal envelope of note: (a)piano, (b)trumpet, (c)violin and (d)organ.

5.3. Partial features

The following features are computed based on locations and amplitudes of the partials obtained in the last section.

- Brightness: the brightness feature B_r is calculated as:

$$B_r = \sum_{k=1}^{N} k a_k / \sum_{k=1}^{N} a_k, \tag{1}$$

  where N is the number of partials, and a_k is the amplitude of the kth partial.

- Tristimulus parameters: the tristimulus features T_{r1}, T_{r2} and T_{r3} as defined below represent energy ratios of the fundamental frequency, the lower to mid range partials, and the partials in mid to higher bands, respectively. In this work, only T_{r1} and T_{r2} are adopted.

$$T_{r1} = a_1 / \sum_{k=1}^{N} a_k, \;\; T_{r2} = (a_2 + a_3 + a_4) / \sum_{k=1}^{N} a_k, \;\; T_{r3} = \sum_{k>4}^{N} a_k / \sum_{k=1}^{N} a_k. \tag{2}$$

- Odd partial ratio: this feature is to detect the lack of energy in odd or even partials. For example, clarinet is well-known for lacking energy in the even partials [10].

$$O_r = \sum_{k=1}^{N/2} a_{2k-1} / \sum_{k=1}^{N} a_k. \tag{3}$$

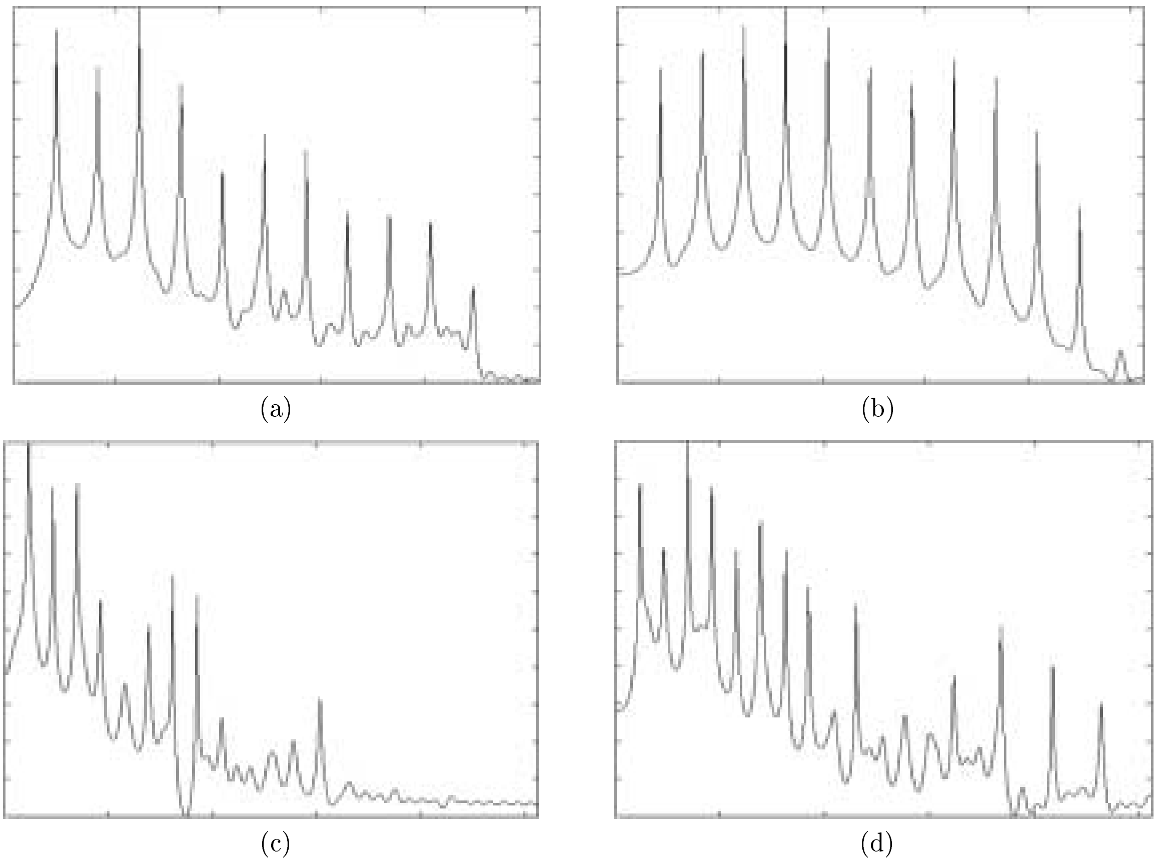


Figure 5. Spectral energy distribution of note: (a)flute, (b)trumpet, (c)guitar and (d)zheng.

- Irregularity: this feature is to measure the amplitude deviation between neighboring partials.

$$I_r = \sum_{k=1}^{N-1} (a_k - a_{k+1})^2 / \sum_{k=1}^{N-1} a_k^2. \tag{4}$$

- Dominant tones: these are the strongest partials in the spectrum. While many instruments generate sounds with strongest partials in lower frequency bands, other instruments produce sounds having strong partials in mid or higher frequency bands. For example, the organ notes have dominant tones at high frequency band as shown in Figure 6(a). Three partials with the highest amplitudes are selected, and their numbers are defined as the dominant tone numbers $D_t = \{D_{t1}, D_{t2}, D_{t3}\}$.

- Inharmonicity: this feature refers to the frequency deviation of partials. As defined, harmonic partials should have frequencies which are integer multiples of the fundamental frequency (FuF). However, some instruments generate sounds which have partials that deviate from the integer multiples of the FuF. A typical example is piano, as illustrated in Figure 6(b) where the vertical lines indicate frequencies which are integer multiples of the FuF. This phenomenon is called inharmonicity. According to [11], the frequency of the partial k is:

$$f_k = k f_0 (1 + I_h \cdot k^2)^{1/2}, \tag{5}$$

where f_0 is the fundamental frequency and I_h is the inharmonicity value. An approximation of the inharmonicity $\hat{I_h}$ is estimated as below:

$$\hat{I_{hi}} = \frac{(\frac{\hat{f_i}}{i\hat{f_0}})^2 - 1}{i^2}, \quad i = 2, \ldots, N, \tag{6}$$

and

$$\hat{I_h} = \frac{\sum_{i=2}^{N} \hat{I_{hi}}}{N-1}, \tag{7}$$

where $\hat{f_i}$ is the estimated frequency of partial i, and $\hat{f_0}$ is the estimated fundamental frequency.

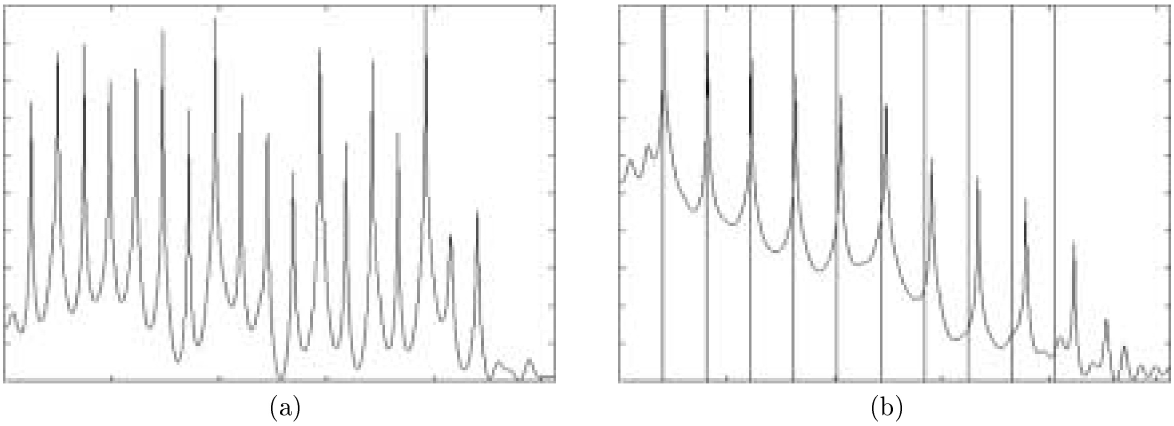


Figure 6. Partial features: (a)dominant tones of an organ note and (b)inharmonicity of a piano note.

5.4. Feature normalization

The normalization process is to ensure that notes of the same instrument have similar feature values and will be classified to the same category, regardless of the loudness or volume, the length and the pitch of the note. The features should also accommodate incomplete notes which may often appear in music pieces, especially in polyphonic music. We also try to keep the value ranges of different features in the same order (i.e. between 0 and 10) which are good for the classification procedures in the following section. The energy of the note has been normalized in Section 4. Here, some of the temporal features and partial features are normalized.

The purpose of normalizing temporal features is to make them independent of the length of the note, and adaptive to incomplete notes. Two features, degree of sustaining D_s and degree of vibration D_v are normalized as follows: take two sets of empirical thresholds, ($D_s$min, $D_s$max) for D_s and ($D_v$min, $D_v$max) for D_v, respectively. Then,

$$\bar{D}_s = \begin{cases} 0 & \text{if } D_s \leq D_s\text{min}, \\ (D_s - D_s\text{min})/(D_s\text{max} - D_s\text{min}) & \text{if } D_s\text{min} < D_s < D_s\text{max}, \\ 1 & \text{if } D_s \geq D_s\text{max}, \end{cases}$$

and

$$\bar{D}_v = \begin{cases} 0 & \text{if } D_v \leq D_v\text{min}, \\ (D_v - D_v\text{min})/(D_v\text{max} - D_v\text{min}) & \text{if } D_v\text{min} < D_v < D_v\text{max}, \\ 1 & \text{if } D_v \geq D_v\text{max}, \end{cases}$$

where $\bar{D}_s$ and $\bar{D}_v$ are the normalized values for D_s and D_v, respectively.

The purpose of normalizing partial features is to make these features independent of the pitch of the note. Three features, the brightness B_r and the tristimulus parameters T_{r1} and T_{r2} are normalized as below:

$$\bar{B}_r = B_r \cdot f_0/1000, \ \ \bar{T}_{r1} = T_{r1} \cdot 1000/f_0, \ \ \bar{T}_{r2} = T_{r2} \cdot 1000/f_0,$$

where f_0 is the estimated value for fundamental frequency, $\bar{B}_r$,$\bar{T}_{r1}$ and $\bar{T}_{r2}$ are the normalized values for B_r,T_{r1} and T_{r2}, respectively. It should be pointed out that the above formula only provides a rough normalization of partial features. For an exact normalization to be done, both the fundamental frequency and the spectral energy distribution pattern should be known *a priori*.

6. NOTE CLASSIFICATION

To classify a note to one kind of instrument, the Kohonen's self-organizing-map (SOM) is employed first to select an optimal structure of the feature vector. The SOM is a type of unsupervised neural network which generates a topological mapping of similarity. That is, similar input samples will correspond to nodes which are close to each other in the 2-D neural net field. Since it is unsupervised neural network, there are no target outputs required for the training samples. However, we have a goal for the overall training process. That is, after training, each kind of instrument should correspond to a region in the neural net field, and similar instruments (e.g. instruments in the same family) should correspond to neighboring regions. For this purpose, the SOM is trained for a number of times, and each time, the training samples have a different feature vector structure, i.e. different selection of features or different ordering of features in the vector. Performances of the obtained SOM neural networks are compared, and the feature vector structure is selected which results in the best performance of the SOM. Feature vectors of all the training and testing samples are rearranged according to this optimal structure.

In the second step, the feature vectors of training samples are sent to a multi-layer-perceptron (MLP), which is a type of supervised neural network having strong classification capabilities. The MLP used in this work has one input layer, one output layer and two hidden layers. The number of nodes at the input layer is the same as the dimension of the note feature vector, and the number of nodes at the output layer is the same as the number of instrument classes. After the MLP is trained, for a given note to be classified, its feature vector is computed and sent as input to the MLP. Suppose there are K classes of instrument, the output of MLP is denoted as $\vec{O} = \{O_1, O_2, \ldots, O_K\}$, where O_i, $i = 1 \sim K$, is the probability of the note's belonging to the ith instrument. Choose the largest element O_m in $\vec{O}$, then, the note is classified to the mth instrument with likelihood O_m.

Since the problem of note classification has fuzzy nature, for example, even a human being may feel ambiguous in recognizing the instrument by hearing only a single note, it is desired that the output of the neural network is also fuzzy. That is, a probability value is more proper rather than an absolute 0 or 1. However, target outputs of available training samples are binary. For instance, the target output of a piano note is a K-dimensional vector in which the element corresponding to piano is 1, and the other elements are 0. To make the MLP have fuzzy features, we propose a two-step training procedure in this work. In the first round of training, target outputs of training samples are binary. After the training process converges, the actual outputs of training samples from the trained MLP are mapped to a predefined distribution, i.e. a linear distribution within a certain range. Then, the mapped outputs are used as target outputs of the training samples for the second round of training.

7. MUSIC PIECE CLASSIFICATION AND RETRIEVAL

As mentioned earlier, one distinction of this work from existing approaches in this area is that we work on continuous music pieces rather than single notes. Compared to single notes, there are more complicated scenarios in music pieces and some useful information for note classification may not be available anymore. For example, partial features at the attacking point of a note may not be available since the attacking point of one note may be overlapped with the ending part of the previous note. On the other hand, one advantage of working on music pieces is that we can integrate features of all the notes within one music piece to get the classification result of the piece. This is consistent to human being's experiences - it is often difficult for us to recognize the instrument by listening to a single note, while it is much easier to distinguish the instrument by listening to a music piece.

For a given music piece to be classified, it is first segmented into notes and each note is categorized to an instrument as described in the above sections. The principle for integrating the note classification results is to count the number of notes classified to each instrument. However, this number is weighted by the likelihood value of each note when it is classified to this instrument. Suppose all the notes in the music piece are grouped into K subsets: $I_1, I_2, \ldots, I_K$, with I_i corresponding to the ith instrument. Then, a score for each instrument is computed as:

$$S(I_i) = \sum_{x \in I_i} O_i(x), \quad i = 1 \sim K, \tag{8}$$

where x denotes a note in the music piece, and $O_i(x)$ is the likelihood of x classified to the ith instrument. After that, the scores are normalized so as to satisfy the following condition:

$$\sum_{i=1}^{K} S(I_i) = 1.$$

Now, take n top scores $S(I_{m1}), \ldots, S(I_{mn})$, with $S(I_{mi}) \geq T_s$, for $i = 1 \sim n$, and $n \leq T_n$. In this work, we choose $T_s = 10\%$ and $T_n = 3$. Again, the selected scores are normalized as:

$$\bar{S}(I_{mi}) = S(I_{mi}) / \sum_{j=1}^{n} S(I_{mj}), \;\; i = 1 \sim n.$$

Finally, the music piece is classified as having instruments numbered $m1$, $m2$, $\ldots$, mn, with percentage $\bar{S}(I_{m1})$, $\bar{S}(I_{m2})$,
$\ldots$, $\bar{S}(I_{mn})$, respectively. Possible results of music piece classification may be:

- The music piece is 100% guitar (with 90% likelihood), or
- The music piece is 60% piano and 40% violin.

Meanwhile, the Gaussian mixture model (GMM) is used in this work for similarity search of music pieces based on timbre features. A GMM is a weighted sum of M component Gaussian densities, where M is selected by the user depending on the complexity of the problem. Its training algorithm is typically an EM process [12]. For each music piece in the database, a GMM is built using feature vectors of all the notes in this music piece. Then, for a query music piece, it is segmented into notes and feature vectors of the notes are computed. Next, these feature vectors are taken to match with GMMs built for the music pieces in the database, and a likelihood is obtained for each GMM. By comparing the likelihood values, a ranking list is generated for the music pieces in terms of similarity with the query piece. And those pieces at top places on the ranking list are regarded as being similar to the query music.

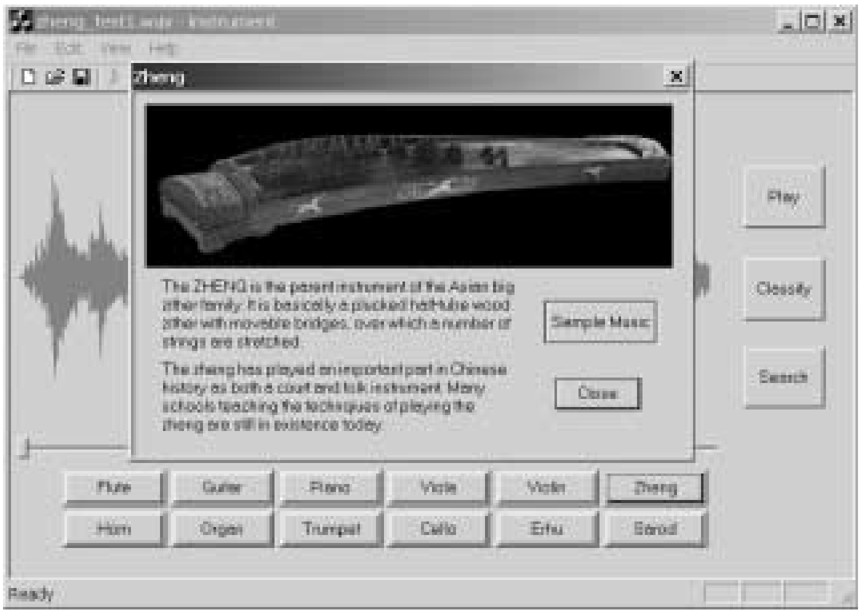
8. EXPERIMENTAL RESULTS

The system has been tested with twelve kinds of musical instruments, including string instruments cello, viola and violin; wind instruments flute, horn and trumpet; together with piano, guitar and organ. Two Chinese instruments erhu (to compare with violin) and zheng (to compare with guitar), and one Indian instrument sarod are also included for comparison purpose. A database containing 287 music pieces was built. These music were extracted from music CDs, including solo and duet music played with one or two instruments, as well as ensemble music having one or two dominant instruments. The length of each music piece is in the range of 10 - 30 seconds. Very promising preliminary results have been obtained through experiments with this database. An accuracy of about 80% is achieved for the identification of dominant instruments in the music pieces. This number can be raised to 90% if classification mistakes within the same instrument family are ignored. For example, cello, viola and violin have certain overlapping ranges, and sometimes it is hard to distinguish among them.

A demonstration system for music instrument classification and music timbre retrieval with user interface design has also been developed, as shown in Figure 7. For the twelve kinds of instrument under consideration in this system, the user can click the button of any instrument and view a picture and a short description of the instrument. A sample music piece of each instrument can also be played. This provides a way for a common user to get familiar with a certain kind of instrument. Illustrated in Figure 8 is a snapshot of the demonstration for instrument recognition. That is, the user can choose an arbitrary music piece in the database, and click the “Classify” button to see which kinds of instruments are involved in this music. Also, by clicking the “Search” button, music pieces in the database which are similar to the selected music in terms of timbre features are retrieved, as shown in Figure 9.

9. CONCLUSION AND FUTURE WORK

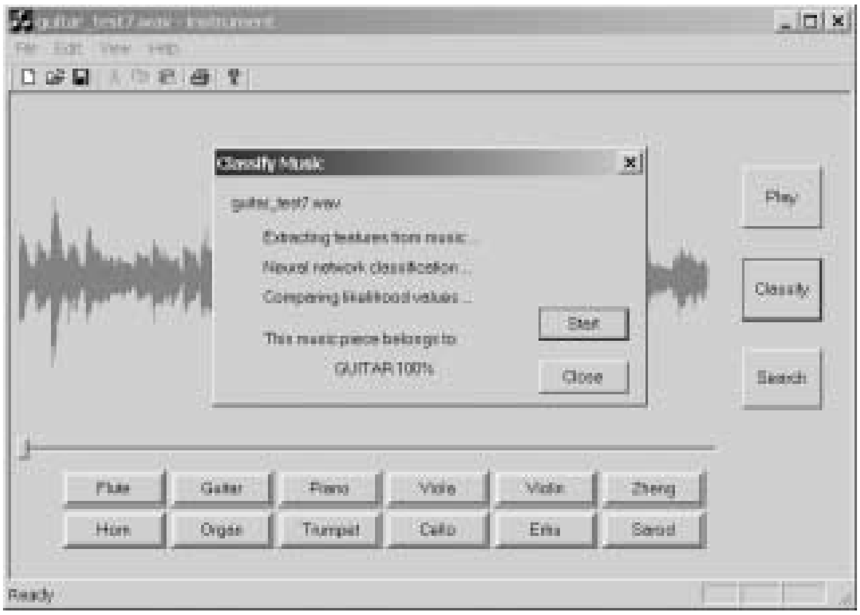
A system for the automatic categorization and retrieval of music assets based on instrument and timbre features was presented in this paper. Compared to previous work in this area, distinguishing features of the proposed system include the following: it has been developed to work with music pieces rather than single notes, to work with real music rather than synthetic music, and to work with both polyphonic and monophonic music. A number of signal processing techniques are integrated in this system, including note onset detection, harmonic partial estimation, note feature extraction and normalization, unsupervised and supervised neural network training, fuzzy logic and so on. The output of the system gives the percentage of each dominant instrument involved in the music piece. Preliminary results obtained through the experiments with twelve kinds of instruments have shown to be quite promising. The developed



The ZHENG is the parent instrument of the Asian big zither family. It is basically a plucked half-tube wood zither with movable bridges, over which a number of strings are stretched.

The zheng has played an important part in Chinese history as both a court and folk instrument. Many schools teaching the techniques of playing the zheng are still in existence today.

Figure 7. User interface of the demonstration.



guitar_test7.wav

Extracting features from music ...

Neural network classification ...

Comparing likelihood values ...

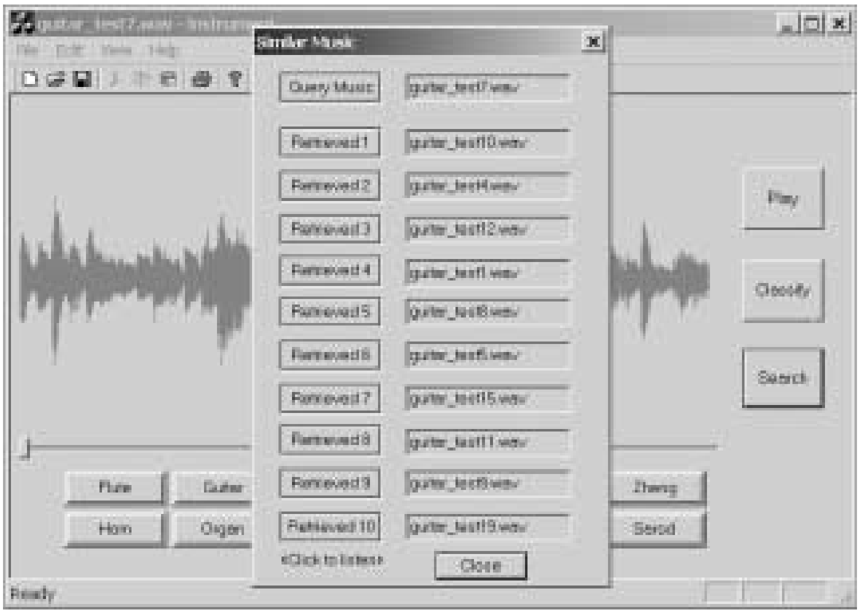
This music piece belongs to:

GUITAR 100%

Figure 8. Demonstration of musical instrument classification.

system may have applications in organizing both professional music databases and personal music collections. It can also be used for online music filtering, browsing and retrieving.

In the next step research, we will first enlarge the music database so that more samples are available to test and fine-tune various modules in this work. The system will also gradually be expanded to include more kinds of instruments. In addition, we are working on defining an optimal set of metadata for the timbre feature of music. For this part, results from some standardization activities, e.g. the MPEG-7 standard which is now being developed, will be referenced. After that, the work will be extended to other music features. Especially, research will be conducted on categorizing and retrieving music based on the rhythm and the melody. Basically, the research area of



Query Music: guitar_test7.wav

Retrieved 1: guitar_test10.wav

Retrieved 2: guitar_test4.wav

Retrieved 3: guitar_test12.wav

Retrieved 4: guitar_test1.wav

Retrieved 5: guitar_test8.wav

Retrieved 6: guitar_test5.wav

Retrieved 7: guitar_test15.wav

Retrieved 8: guitar_test11.wav

Retrieved 9: guitar_test9.wav

Retrieved 10: guitar_test19.wav

Figure 9. Demonstration of music similarity search based on timbre features.

music database management is still quite blank, and there is huge amount of work left to be done. While there are numerous challenging problems to be solved, some promising results have been achieved.

REFERENCES

1. B. Kostek, A. Czyzewski, "Automatic classification of musical timbre based on learning algorithms applicable to cochlear implants", *Proceedings of Artificial Intelligence, Expert Systems and Neural Networks*, pp.98-101, 1996.
2. A. Wieczorkowska, "Rough sets as a tool for audio signal classification", *Proceedings of International Symposium on Methodologies for Intelligent Systems*, pp.367-75, June 1999.
3. J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features", *Journal of the Acoustical Society of America*, vol.105, no.3, pp.1933-41, 1999.
4. J. Jeong, D. Kim, S. Lee, "Musical timbre recognition with neural networks", *Proceedings of International Conference on Neural Information Processing*, vol.2, pp.869-72, Oct. 1995.
5. P. Cosi, G. De Poli, G. Lauzzana, "Auditory modeling and self-organizing neural networks for timbre classification", *Journal of New Music Research*, vol.23, no.1, pp.71-98, 1994.
6. S. Sayegh, C. Pomalaza-Raez, M. Badie, *et al.*, "A neural network approach to timbre discrimination of identical pitch signals", *Journal of Intelligent Systems*, vol.7, no.3-4, pp.339-47, 1997.
7. D. K. Fragoulis, J. N. Avaritsiotis, C. N. Papaodysseus, "Timbre recognition of single notes using an ARTMAP neural network", *Proceedings of IEEE International Conference on Electronics, Circuits and Systems*, vol.2, pp.1009-12, Sep. 1999.
8. G. Costantini, F. M. Frattale Mascioli, A. Rizzi, *et al.*, "Recognition of musical instruments by a nonexclusive neuro-fuzzy classifier", *Proceedings of Second EURASIP Conference on DSP for Multimedia Communications and Services*, June 1999.
9. J. Mariano Merino, "Complexity of pitch and timbre concepts", *Physics Education*, vol.33, no.2, pp.105-9, 1998.
10. N. H. Fletcher, T. D. Rossing, *The Physics of Musical Instruments*, Springer-Verlag, 1993.
11. K. Jensen, "Spectral envelope modeling", DSAGM, DIKU, Aug. 1998.
12. D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture models", *IEEE Transactions on Speech and Audio Processing*, vol.3, no.1, pp.72-83, 1995.