Analysis of Sound Features for Music Timbre Recognition

Xin Zhang and Zbigniew W. Ras
Department of Computer Science
University of North Carolina at Charlotte
xinzhang@uncc.edu, ras@uncc.edu

Abstract

Recently, communication, digital music creation, and computer storage technology has led to the dynamic increasing of online music repositories in both number and size, where automatic content-based indexing is critical for users to identify possible favorite music pieces. Timbre recognition is one of the important subtasks for such an indexing purpose. Lots of research has been carried out in exploring new sound features to describe the characteristics of a musical sound. The Moving Picture Expert Group (MPEG) provides a standard set of multimedia features, including low level acoustical features based on latest research in this area. This paper introduces our newly designed temporal features used for automatic indexing of musical sounds and evaluates them with MPEG7 descriptors, and other popular features.

1. Introduction

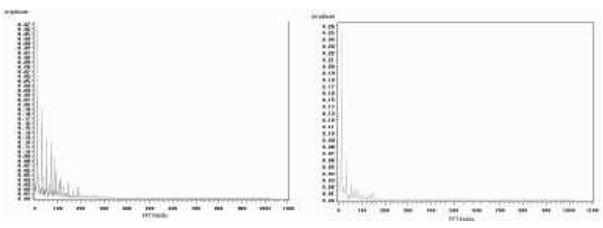
In recent years, researchers have extensively investigated lots of acoustical features to build computational model for automatic music timbre estimation. Timbre is a quality of sound that distinguishes one music instrument from another, while there are a wide variety of instrument families and individual categories. It is rather 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 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 is very nicely discussed in [3]. The following are some of the specific challenges that motivated the development of new sound features in this paper.

Enormous data size - a digital musical object may consist of lots of subtle changes, which is noticeable or even critical to human sound perception system.

High dimensionality - most western orchestral instruments have rich timbre and produce overtones, which results in a sound with a group of frequencies in clear mathematical relationships (so-called harmonics). There are many different approaches to detect sound timbre (for instance [2] or [4]). Some of them are quite successful on certain simply sound data (monophonic, short, of limited instrument types). Dimensional approach to timbre description was proposed in [3]. Timbre description is basically subjective and vague, and only some subjective features have well defined objective counterparts, like brightness, calculated as gravity center of the spectrum. Explicit formulation of rules of objective specification of timbre in terms of digital descriptors will formally express subjective and informal sound characteristics. It is especially important in the light of human perception of sound timbre. Therefore, evolution of sound features in time should be reflected in sound description as well. The discovered temporal patterns may better express sound features than static features, especially that classic features can be very similar for sounds representing the same family or pitch, whereas changeability of features with pitch for the same instrument makes sounds of one instrument dissimilar. Therefore, classical sound features can make correct identification of musical instrument independently on the pitch very difficult and erroneous.

Methods in research on automatic musical instrument sound classification go back to last few years. So far, there is no standard parameterization used as a classification basis. The sound descriptors used are based on various methods of analysis in time

domain, spectrum domain, time-frequency domain and cepstrum with Fourier Transform for spectral analysis being most common, such as Fast Fourier Transform (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, in which it described a set of low-level sound temporal and spectral features. The low-level descriptors in MPEG-7 are intended to describe the time-variant information within an entire audio segment, where most of them are, like other STFT related acoustic features, in a form of either vector or matrix of large size, where an audio segment was divided into a set of frames and each row represents a power spectrum in the frequency domain within each analysis window. Therefore, these features are not suitable for traditional classifiers, which require single-value cell of input datasets. Researchers have been explored different statistical summations in a form of single value to describe signatures of music instruments within vectors or matrices in those features, such as Tristimulus parameters [17], Brightness [7], and Irregularity [21], etc. However, current features fail to sufficiently describe the audio signatures which vary in time within a whole sound segment, esp. where multiple audio signatures are overlapping with each other. It was widely observed that a sound segment of a note, which is played by a music instrument, has at least three states: transient state, quasi-steady state and decay state. Vibration pattern in a transient state is known to significantly differ from the one in a quasi-steady state. Consequently, the harmonic features in the transient state behavior are significantly different from those in the quasi-steady state. In the figure below, the left graph shows the power spectrum in linear scale in the transient state of 3A flat clarinet (a monophonic sound), where energy is distributed around a few harmonic peaks; the right graph shows the power spectrum in the quasi-steady state of the same sound, where the energy is more evenly distributed around several harmonic peaks.



Time-variant information is necessary for correct classification of musical instrument sounds, because quasi-steady state itself is not sufficient for human experts. Also, it has been observed that a human needs the beginning of the music sound to discern the type of an instrument. Identifying the boundary of the transient state enables accurate timbre recognition. Wieczorkowska et. al [22] proposed a timbre detection system with differentiated analysis in time, where each sound segment has been split into seven intervals of equal width. However, the length of the duration of transient state varies from one instrument to another, thus it is difficult to find a universal quantization approach with fixed number of bins for sounds of all instruments. In our research, we have proposed new approach to differentiate the states of a segment for harmonic feature analysis.

2. Audio Features in our research

There are many ways to categorize the audio features. In this paper, audio features in our system 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, a derivative database of those features with single valued data for KDD classification will be 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.

2.1. MPEG7 based descriptors

Based on latest research in the area, MPEG published a standard of a 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.

$$P_x(k), k = 0,..,\frac{NFFT}{2} \quad 1.),$$

$$C = \sum_n \log_2(f(n)/1000) P_x'(n) \Big/ \sum_n P_x'(n) \quad 2.),$$

where *Sr* is the sample rate. 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.

$$S = \sqrt{\sum_n ((\log_2(f(n)/1000) - C)^2 P_x'(n)) \Big/ \sum_n P_x'(n)} \quad 3.),$$

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, which is ranged by edges in the following formula.

$$edge = 2^{0.25m} \times 1KHz \quad 4.),$$

$$SFM_b = \frac{\sqrt[ih(b)-il(b)+1]{\prod_{i=il(b)}^{ih(b)} c(i)}}{\frac{1}{ih(b)-il(b)+1}\sum_{i=il(b)}^{ih(b)} c(i)} \quad 5.),$$

where $c(i)$ is the mean value of a group of power spectrum coefficients, and the total number of each group is determined by the location of each frequency bin. The value of each bin is treated as an attribute value in the database. Since the octave resolution in the thesis is 1/4, the total number of bands is 32.

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. x_t is a vector of power spectrum coefficients in a frame t, which are transformed to Db scale and then normalized. N, the total number of frequency bins, is 32 in 1/4 octave resolution.

$$\chi = 10\log_{10}(x_t) \quad 6.),$$

$$r = \sqrt{\sum_{k=1}^{N} \chi_k^2} \quad 7.),$$

$$\tilde{X} = \frac{\chi}{r} \quad 8.),$$

$$\tilde{\tilde{X}} = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \vdots \\ \tilde{x}_M^T \end{bmatrix} \quad 9.),$$

$$\tilde{\tilde{X}} = USV^T \quad 10.),$$

$$V_K = [v_1 \quad v_2 \quad \cdots \quad v_k] \quad 11.),$$

Spectrum Projection Functions

$$y_t = [r_t \quad \tilde{x}_t^T v_1 \quad \tilde{x}_t^T v_2 \quad \cdots \quad \tilde{x}_t^T v_k] \quad 12.),$$

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.

$$IHSC(frame) = \frac{\sum_{harmo=1}^{nb\_harmo} f(frame, harmo) \cdot A(frame, harmo)}{\sum_{harmo=1}^{nb\_harmo} A(frame, harmo)} \quad 13.),$$

$$HSC = \frac{\sum_{frame=1}^{nb\_frames} IHSC(frame)}{nb\_frames} \quad 14.),$$

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.

$$IHSS(i) = \frac{1}{IHSC(i)} \sqrt{\frac{\sum_{k=1}^{K} A^2(i,k) \cdot [f(i,k) - IHSC(i)]^2}{\sum_{k=1}^{K} A^2(i,k)}} \quad 15.),$$

$$HSS = \frac{\sum_{i=1}^{M} IHSS(i)}{M} \quad 16.),$$

where A is the power of the k_{th} harmonic peak in the i_{th} frame, K is the total number of harmonic peaks, M is the total number of frames in a music object.

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.

$$IHSV(i) = 1 - \frac{\sum_{k=1}^{K} A(i-1,k) \cdot A(i,k)}{\sqrt{\sum_{k=1}^{K} A^2(i-1,k)} \cdot \sqrt{\sum_{k=1}^{K} A^2(i,k)}} \quad 17.),$$

$$HSV = \frac{\sum_{k=1}^{K} IHSV(i)}{M} \quad 18.),$$

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.

$$SE(i,k) = \frac{A(i,k) + A(i,k+1)}{2} \quad 19.),$$

$$SE(i,k) = \frac{\sum_{j=-1}^{1} A(i,k+j)}{3}, k = 2, K-1 \quad 20.),$$

$$IHSD(i) = \frac{\sum_{k=1}^{K} |\log_{10}(A(i,k)) - \log_{10}(SE(i,k))|}{\sum_{k=1}^{K} \log_{10}(A(i,k))} \quad 21.),$$

$$HSD = \frac{\sum_{i=1}^{M} IHSD(i)}{M} \quad 22.),$$

where A stands for amplitude of a harmonic peak in a frame.

Log Attack Time is defined as the logarithm of the time duration between the time 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.

$$LAT = \log_{10}(T1 - T0) \quad 23.),$$

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.

$$r(i,k) = \frac{\sum_{j=m}^{m+n-1} s(j)s(j-k)}{\left(\sum_{j=m}^{m+n-1} s(j)^2 \times \sum_{j=m}^{m+n-1} s(j-k)^2 \right)^{0.5}} \quad 24.),$$

$$H(i) = \max_{k=Q}^{n-1} r(i,k) \quad 25.),$$

Fundamental Frequency is the frequency that best explains the periodicity of a signal. The ANSI definition of psycho-acoustical terminology says that "pitch is that auditory attribute of sound according to which sounds can be ordered on a scale from low to high". It is estimated based on the local maximums of the $r(i,k)$, which is normally in shape of a sinusoid with amplitudes ranging from –1 to 1. The figure below shows the first, fourth, fifth, and fourteenth frames of a sound in the first octave F in the order of a, b, c, d, which was played by an electric bass. This pattern varies from frame to frame, especially where a sound state is changed. In some frames, the range of the correlation function value is from 0 to 1 as shown in pattern c; in other frames, there are complex sinusoid patterns, where each periodical consists of a set of sub-peaks anchoring either around 0 or 0.5. Therefore, zero crossing is not suitable to search for local peaks.

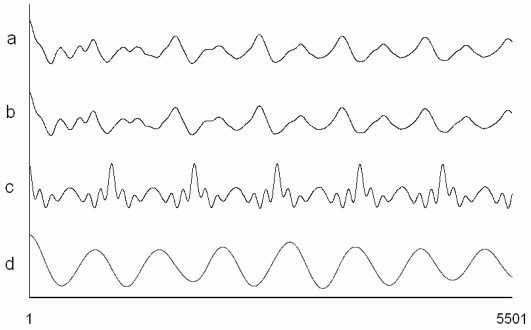


Figure 1. Cross-correlation pattern for pitch estimation.

Normally the first few points have highest values, especially when the pitch is very low, thus the lag values at the beginning are negligible comparing to the long periodicity. Therefore, this part of lags should be skipped while searching for the maximum peak. The starting lag Q is calculated by this formula:

$$r(i,k) < c * r(i,n), n = Sr / f0' \qquad 26.),$$

where k is the maximum position of the lag, at which $r(i,k)$ is less than a flexible threshold according to the first lag position, c is an empirical threshold (0.2), $f0'$ is the expected maximum fundamental frequency.

MPEG7 suggests take the first maximum in order to estimate the local fundamental frequency period. Since the energy of local peaks in the center of the pattern normally presents a more stable periodicity character than that of the ones at the beginning or the end, we adapt this method by taking the difference between the maximum peak and the immediate previous local peak of it. The instantaneous fundamental frequency is then estimated by the inverse of the time corresponding to the difference of those two positions. We observed significant improvement of the performance and the accuracy, especially for the low frequency sounds where MPEG7 algorithm fails.

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.

$$c(j) = s(j) - \lambda s(j-K),\ j = m,(m+n-1) \qquad 27.),$$

$$\lambda = \sum_{j=m}^{m+n-1} s(j)s(j-K) \Big/ \sum_{j=m}^{m+n-1} s^2(j-K) \qquad 28.),$$

$$a(f_{\lim}) = \sum_{f=f_{\lim}}^{f_{\max}} p'(f) \Big/ \sum_{f=f_{\lim}}^{f_{\max}} p(f) \qquad 29.),$$

$$ULH(i) = \log_2(f_{u\lim}/1000) \qquad 30.),$$

where *c(j)* is a comb-filtered sample data, *K* is the lag corresponding to the maximum cross correlation *H(i)*, *p(f)* and *p'(f)* are the power spectrum coefficients of the original signal and the combed signal in the *ith* frame.

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.

$$S(k) = \sqrt{\sum_{i=1}^{M} P_i(k) \Big/ M} \qquad 31.),$$

$$SC = \frac{\sum_{k=1}^{NFFT} f(k) \cdot S(k)}{\sum_{k=1}^{NFFT} S(k)} \qquad 32.),$$

where M is the total number of frames in a sound segment, $P_i(k)$ is the *kth* power spectrum coefficient in the *ith* frame, f(k) is the *kth* frequency bin.

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

$$TC = \frac{\sum_{n=1}^{length(SEnv)} n/sr \cdot SEnv(n)}{\sum_{n=1}^{length(SEnv)} SEnv(n)} \quad 33.),$$

2.2. Other descriptors

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

Vector descriptors. Since $\mathbf{V}_K$ is 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 and row, where the dissimilarity is measured by the following equation:

Tristimulus parameters describe the ratio of the amplitude of a harmonic partial to the total harmonic partials [17]. 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.

$$Tr_1 = A_1^2 \Big/ \sum_{n=1}^{N} A_n^2 \quad 34.),$$

$$h_{3,4} = \sum_{i=3}^{4} A_i^2 \Big/ \sum_{j=1}^{N} A_j^2 \quad 35.),$$

$$h_{5,6,7} = \sum_{i=5}^{7} A_i^2 \Big/ \sum_{j=1}^{N} A_j^2 \quad 36.),$$

$$h_{8,9,10} = \sum_{i=8}^{10} A_i^2 \Big/ \sum_{j=1}^{N} A_j^2 \quad 37.),$$

$$Od = \sqrt{\sum_{k=2}^{L} A_{2k-1}^2} \Big/ \sqrt{\sum_{n=1}^{N} A_n^2} \quad 38.),$$

$$Ev = \sqrt{\sum_{k=1}^{M} A_{2k}^2} \Big/ \sqrt{\sum_{n=1}^{N} A_n^2} \quad 39.),$$

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

$$B = \sum_{n=1}^{N} n \cdot A_n \Big/ \sum_{n=1}^{N} A_n \quad 40.),$$

$$\overline{fd}_m = \sum_{k=1}^{5} A_k (\Delta f_k / (k f_1)) / \sum_{k=1}^{5} A_k \quad 41.),$$

Transient, steady and decay duration. In this research, the transient duration is considered as the time to reach the quasi-steady state of fundamental frequency. 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.

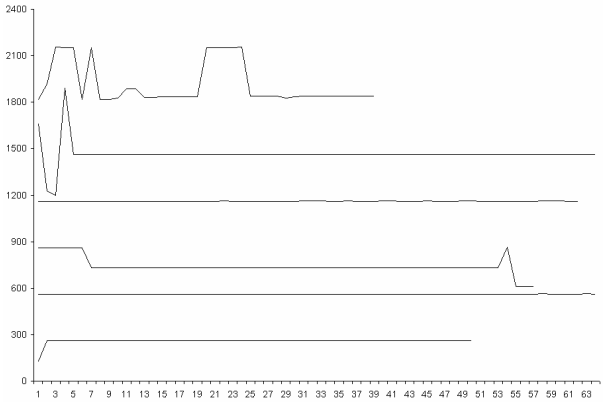


Figure 2. Pitch trajectories of note 4C played by different instruments.

We observe that during the transients, the instantaneous fundamental frequencies are unstable, and usually very different from the ones in the quasi-steady state, see above figure. The transient border is estimated as the first frame where the pitch stays considerately stable during a minimum period of time. It is computed as the total number of the continuous frames with similar instantaneous fundamental

frequencies, which is bigger than a time-duration threshold. Due to the wide range of the length of the sample recordings, which is from around 26 to over 300 frames, and the fact that short sounds are, in most cases, short in each state, three different empirical threshold values of time duration are applied according to the total length of each music object. For objects less than 30 frames, the threshold was set to three, which was a 30 milliseconds and was more than 10% of its total length; for objects less than 100 and longer than 30 frames, the threshold was set to five, which was 70 milliseconds and was more than 5% of its total length; for objects longer than 100 frames, the threshold was set to eight, which was 100ms.

The beginning of the quasi-steady state is at the first frame having an overall fundamental frequency in the same frequency bin *as its N* continuous following neighbor frames, where the total energy in the spectrum is bigger than a threshold in case of salience or noise. Each frequency bin corresponds to a music note. The overall fundamental frequency is estimated by pattern recognition with a cross-correlation function.

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 [19] [20].

$$ZC_i = 0.5\sum_{n=1}^{N}\left|sign(s_i[n]) - sign(s_i[n-1])\right| \qquad 42.),$$

$$sign(x) = \begin{cases} 1, x \geq 0 \\ -1, x < 0 \end{cases} \qquad 43.),$$

where s_i is the n^{th} sample in the i^{th} frame, N is the frame size.

Spectrum Centroid describes the gravity center of the spectrum [19] [23].

$$C_i = \frac{\sum_{k=1}^{N/2} f(k)\left|X_i(k)\right|}{\sum_{k=1}^{N/2}\left|X_i(k)\right|} \qquad 44.),$$

where N is the total number of the FFT points, $X_i(k)$ is the power of the kth FFT point in the ith frame, f(k) is the corresponding frequency of the FFT point.

Roll-off is a measure of spectral shape, which is used to distinguish between voiced and unvoiced speech [13]. 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.

$$\sum_{k=1}^{K}\left|X_i(k)\right| \leq C \cdot \sum_{k=1}^{K}\left|X_i(k)\right| \qquad 45.),$$

Flux is used to describe the spectral rate of change [19]. It is computed by the total difference between the magnitude of the FFT points in a frame and its successive frame.

$$F_i = \sum_{k=1}^{N/2}\left(\left|X_i(k)\right| - \left|X_{i-1}(k)\right|\right)^2 \qquad 46.),$$

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).

3. Classification

The classifiers, applied in the investigations on musical instrument recognition, represent practically all known methods. In our research, so far we have used four classifiers (Bayesian Networks and Decision Tree J-48) upon numerous music sound objects to explore the effectiveness of our new descriptors.

Naïve Bayesian is a widely used statistical approach, which represents the dependence structure between multiple variables by a specific type of graphical model, where probabilities and conditional-independence statements are strictly defined. It has been successfully applied to speech recognition [26] [14].

The joint probability distribution function is

$$p(x_1,...,x_N) = \prod_{i=1}^{N} p(x_i \mid x_{\pi i}) \qquad 47.),$$

where x_1 , …, x_N are conditional independent variables, and $x_{\pi i}$ is the parent of x_i .

Decision Tree-J48 is a supervised classification algorithm, which has been extensively used for machine learning and pattern recognition. See [18] [21]. 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.

4. Experiments

We used a database of 3294 music recording sound objects of 109 instruments from McGill University Master Samples CD Collection, which has been widely used for research on musical instrument recognition all over the world. We implemented a hierarchical structure in which we first discriminate different instrument families (woodwind, string family, harmonic percussion family and non-harmonic family), and then discriminate the sounds into different type of instruments within each family. The woodwind family had 22 different instruments. The string family contained seven different instruments. The harmonic percussion family included nine instruments. The non-harmonic family contained 73 different instruments. All classifiers were 10-fold cross validation with a split of 90% training and 10% testing. We used WEKA for all classifications.

In this research, we compared hierarchical classifications with none- hierarchical classifications in our experiments. The hierarchical classification schema is shown in the following figure.

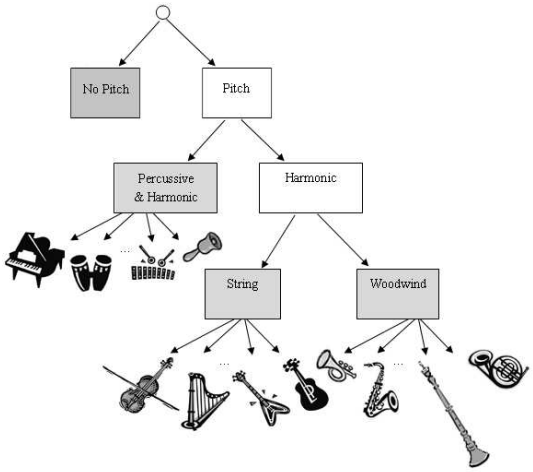


Figure 3. Hierarchical Classification.

The audio objects in our database are grouped into four different types: sounds without pitch, percussive and harmonic sounds, sounds played by the string instruments, and sound played by the woodwind instruments. Each type contains all different instruments.

TABLE 1. Results of the hierarchical Classification with all the features.

	J48-Tree	NaïveBaysian
Family	91.726%	72.6868%
No-pitch	77.943%	75.2169%
Percussion	86.0465%	88.3721%
String	76.669%	66.6021%
Woodwind	75.761%	78.0158%

Table1 shows the performance of the classifiers constructed with all the features. Family represents classifications to distinguish the instrument family type. No-pitch represents classifications of all the audio objects, which are belonging to the non-harmonic family. Percussion stands for harmonic percussion instruments.

TABLE 2. Results of the hierarchical Classification with the MPEG7 features.

	J48-Tree	NaïveBaysian
Family	86.434%	64.7041%
No-pitch	73.7299%	66.2949%
Percussion	85.2484%	84.9379%
String	72.4272%	61.8447%
Woodwind	67.8133%	67.8133%

Table2 shows the performance of the classifiers constructed with only the MPEG7 features.

TABLE 3. Results of the none-hierarchical Classification.

	J48-Tree	NaïveBaysian
All	70.4923%	68.5647%
MPEG	65.7256%	56.9824%

Table3 shows the performance of the classifiers constructed without hierarchical schema.

5. Conclusion and future work

By the results from the experiments, we conclude that the non-MPEG features significantly improved the performance of the classifiers. However, for sounds played by instruments from the string family and the woodwind family, there is no significant improvement by adding the hierarchical schema into the system. Since we consider the transient duration as the time to reach the stable state of the pitch of a note, the transient state contains more information, which is highly correlated to the timbre properties and less relevant to the pitch properties. For classification of the instrument families, the feature vectors from the transient state have a better overall performance than those from the quasi-steady state and from the whole segment over most classifiers except for Decision Tree J48. Similar results we observed on identification of the instruments within the String family. We also observed that the Woodwind family has shorter transient duration than the String family in average. The frame size and hop size need to be adjusted to capture subtle variation in the transient duration. It could explain the fact that feature vector from the quasi-steady state have better overall performance than those from the transient state and the whole segment for classification of the instruments within the Woodwind family.

The proposed research is still in its early stages. More new features in different acoustic states and music objects from more instruments families in different articulations will be investigated. Also, future research shall explore the efficiency of the classification system based on segmented sequence of music piece.

6. Acknowledgment

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

7. References

[1] Atkeson, C.G., Moore A.W., and Schaal, S. (1997). Locally Weighted Learning for Control, Artificial Intelligence Review. Feb. 11(1-5), 75-113.

[2] Balzano, G.J. (1986). What are musical pitch and timbre? Music Perception - an interdisciplinary Journal. 3, 297-314.

[3] Bregman, A.S. (1990). Auditory scene analysis, the perceptual organization of sound, MIT Press

[4] Cadoz, C. (1985). Timbre et causalite, Unpublished paper, Seminar on Timbre, Institute de Recherche et Coordination Acoustique / Musique, Paris, France, April 13-17.

[5] Dziubinski, M., Dalka, P. and Kostek, B. (2005) Estimation of Musical Sound Separation Algorithm Effectiveness Employing Neural Networks, Journal of Intelligent Information Systems, 24(2/3), 133–158.

[6] Eronen, A. and Klapuri, A. (2000). Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. In proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, Plymouth, MA, 753-756.

[7] Fujinaga, I., McMillan, K. (2000) Real time Recognition of Orchestral Instruments, International Computer Music Conference, 141-143.

[8] Gillet, O. and Richard, G. (2005) Drum Loops Retrieval from Spoken Queries, Journal of Intelligent Information Systems, 24(2/3), 159-177

[9] Herrera. P., Peeters, G., Dubnov, S. (2003) Automatic Classification of Musical Instrument Sounds, Journal of New Music Research, 32(19), 3–21.

[10] Kaminskyj, I., Materka, A. (1995) Automatic source identification of monophonic musical instrument sounds, the IEEE International Conference On Neural Networks.

[11] Kostek, B. and Wieczorkowska, A. (1997). Parametric Representation of Musical Sounds, Archive of Acoustics, Institute of Fundamental Technological Research, Warsaw, Poland, 22(1), 3-26.

[12] le Cessie, S. and van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression, Applied Statistics, 41, (1), 191-201.

[13] 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.

[14] Livescu, K., Glass, J., and Bilmes, J. (2003). Hidden Feature Models for Speech Recognition Using Dynamic Bayesian Networks, in Proc. Euro-speech, Geneva, Switzerland, September, 2529-2532.

[15] Logan, B. Mel (2000). Frequency Cepstral Coefficients for Music Modeling, in Proc. 1st Ann. Int. Symposium On Music Information Retrieval (ISMIR).

[16] 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.

[17] Pollard, H.F. and Jansson, E.V. (1982). A tristimulus Method for the spectificaiton of Musical Timbre. Acustica, 51, 162-171

[18] Quinlan, J.R. (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA.

[19] 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).

[20] Tzanetakis, G. and Cook, P. (2002)"Musical Genre Classification of Audio Signals," IEEE Trans. Speech and Audio Processing, July, vol. 10, pp. 293–302.

[21] 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.

[22] Wieczorkowska, A., Wroblewski, J., Synak, P., and Slezak, D. (2003). Application of Temporal Descriptors to Musical Instrument Sound, Journal of Intelligent Information Systems, Integrating Artificial Intelligence and Database Technologies, July, 21(1), 71-93.

[23] Wold, E., Blum, T., Keislar, D., and Wheaton, J.,(1996). Content-Based Classification, Search and Retrieval of Audio, IEEE Multimedia,Fall, pp. 27–36.

[24] Zhang, X. and Ras, Z.W. (2006A). Differentiated Harmonic Feature Analysis on Music Information Retrieval For Instrument Recognition, proceeding of IEEE International Conference on Granular Computing, May 10-12, Atlanta, Georgia, 578-581.

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

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

[27] ISO/IEC JTC1/SC29/WG11 (2002). MPEG-7 Overview. http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm