Sound Isolation by Harmonic Peak Partition For Music Instrument Recognition

Xin Zhang
Department of Computer Science,
University of North Carolina, Charlotte
NC 28223, USA
xinzhang@uncc.edu

Zbigniew W. Raś
Department of Computer Science
University of North Carolina, Charlotte
NC 28223, USA
ras@uncc.edu

Abstract. Identification of music instruments in polyphonic sounds is difficult and challenging, especially where heterogeneous harmonic partials are overlapping with each other. This has stimulated the research on sound separation for content-based automatic music information retrieval. Numerous successful approaches on musical data feature extraction and selection have been proposed for instrument recognition in monophonic sounds. Unfortunately, none of those algorithms can be successfully applied to polyphonic sounds. Based on recent successful researches in sound classification of monophonic sounds and studies in speech recognition, Moving Picture Experts Group (MPEG) standardized a set of features of the digital audio content data for the purpose of interpretation of the information meaning. Most of them are in a form of large matrix or vector of large size, which are not suitable for traditional data mining algorithms; while other features in smaller size are not sufficient for instrument recognition in polyphonic sounds. Therefore, these acoustical features themselves alone cannot be successfully applied to classification of polyphonic sounds. However, these features contains critical information, which implies music instruments' signatures. We proposed a novel music information retrieval system with MPEG-7-based descriptors and we built classifiers which can retrieve the important time-frequency timbre information and isolate sound sources in polyphonic musical objects, where two instruments are playing at the same time, by energy clustering between heterogeneous harmonic peaks.

Keywords: Music Instruments Detection, MPEG-7 descriptors, Musical Sound Separation, Energy

Clustering, Sound Classification, and Feature Extraction

1. Introduction

There are a wide variety of applications in Automatic Music Information Retrieval, where purposes are as different, as medical diagnosis can be from surveillance or military operations. Recently, the booming of multimedia resources on the Internet brought a tremendous need to provide new, more advanced tools for the ability to query and process vast quantities of musical data, which are not easy to describe with mere symbols. However, searching through multimedia data is a highly nontrivial task that requires content-based indexing of the data. Lots of multimedia-resources provide data that have been manually labeled with some description information, such as title, author, company, and so on. However, in most cases those labels are insufficient for content-based searching. Generally, identification of musical information can be performed for digital audio data, like audio samples taken from real recordings, representing waveform, and for MIDI (Musical Instrument Digital Interface) data, etc. MIDI files give access to highly structured data. So, research on MIDI data may basically concentrate on higher level of musical structure, like key or metrical information. Identifying the dominant instruments, which are playing in the multimedia segments, is even more difficult. Timbre 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 [4].

Most western orchestral instruments have rich timbre and produce overtones, which result in a sound with a group of frequencies in clear mathematical relationships (so-called harmonics). There are a number of different approaches to detect sound timbre (for instance [3] or [7]). Some of them are quite successful on certain simply sound data (monophonic, short, of limited instrument types). Dimensional approach to timbre description was proposed by [4]. 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 [19], Brightness [10], and Irregularity [24], etc. However, current features can not sufficiently describe the audio signatures which vary in time within a whole sound segment, esp. where multiple audio signatures are overlapping with each other. This has stimulated differentiated studies in the time-frequency domain and research on sound separation.

Over the last twenty years, researchers on Blind Source Separation (BBS) and independent component analysis (ICA) have been extensively explored numerous approaches to isolate sounds within sound mixing from multiple sound sources. BBS is to estimate original sound sources based on signal observations without any knowledge on the mixing and filter procedure. ICA is to separate sounds by linear models of matrix factorization based on the assumption that each sound source is statistically independent. Applications of BBS by ICA are mostly in medical signal processing and speech recognition [21], [2], [22]. However, due to the lack of precise models for harmonic patterns and behaviors, estimation of original sounds has been difficult and erroneous, since the estimation of the component distributions can be sometimes erroneous or even fatal. Another importance trend for BBS is to use sinusoidal modeling techniques and filter bank to separate harmonic sounds. Based on the fact that harmonic components have significant energy, harmonics tracking with Q-Constant Transform and Short Time Fourier Transform has been applied to sound separation [16], [12]. In this paper, multi-pitch recognition will not be discussed. To focus only on harmonic partial clustering, we assume that multi-pitch values of polyphonic sounds are known.

It was widely observed that a sound segment of a note, which is played by a music instrument, have at least three states: transient state, quasi-steady state and decay state. Vibration pattern in a transient state is known to significantly differ from the one in a quasi-steady state. Consequently, the harmonic features in the transient state behavior significantly different from those in the quasi-steady state. See figure (3),(6).

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. [24] proposed a timbre detection system with differentiated analysis in time, where each sound segment has been split into 7 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.

Popular classifiers, which have been applied to automatic musical sound classification and speech recognition, include the k-Nearest Neighbors algorithm (k-NN see [15] and [10]), Naive Bayesian Clas-

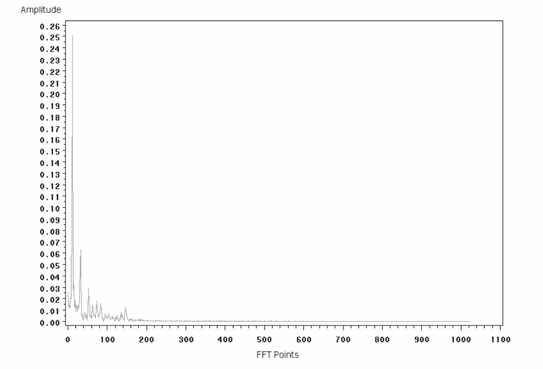

Figure 1. The power spectrum in the transient state of a of 3A B-flat clarinet (monophonic sound). Amplitude in linear scale. Energy is distributed around a few harmonic peaks.

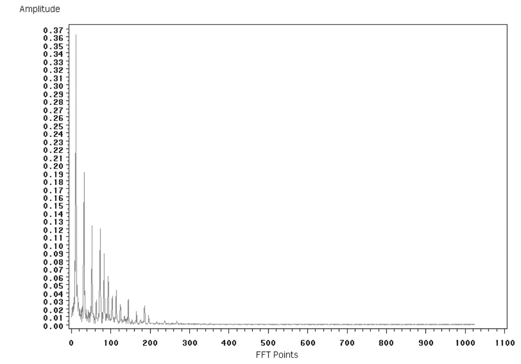

Figure 2. The power spectrum in the quasi-steady state of a of 3A B-flat clarinet (monophonic sound). Amplitude in linear scale. Energy is more evenly distributed around several harmonic peaks.

sifiers (see [17], [5]), Decision Trees (see [13]), Discriminant Analysis, Higher Order Statistics, Artificial Neural Networks, Support Vector Machines, Rough Sets and Hidden Markov Models, etc. Hierarchical classification procedures together the k-NN algorithm and other classifiers have been explored in different sound recognition systems (see [18] [9]), and were reported as classifiers which increased the classification accuracy in comparison to the non-hierarchical classification systems. However, due to the wide variety of feature behaviors of different musical sounds, no classifier has been reported as significantly better than others in all cases. Thus we have chosen four different classifiers to classify the separated musical sounds in our research.

2. Sound Isolation Algorithm

Harmonics are integer multiples of fundamental frequency. In a power spectrum from STFT or other discrete fourier transform, energies are distributed closely on several FFT points near each harmonic peak. In a polyphonic sound with multiple pitches, multiple sets of harmonics from different instrument sources are overlapping with each other. For example, in a sound mix where a sound in 3A of clarinet and a sound in 4C of violin were played at the same time, there are two sets of harmonics: one set is distributed near several integer multiples of 440Hz; the other spreads around integer multiples of 523.25Hz. In this research, we applied a clustering technique to separate the energy of FFT points in the power spectrum by taking each harmonic peak as a cluster centroid. Therefore, sound separation has been implemented by clustering the energy around each harmonic peak. See figure (4),(5),(3).

We processed audio sample segments by a discrete STFT into w sub-segments, which were overlapping with each other in time, with a Hamming window to prevent stray. The hop size was set to 10 milliseconds, which is within the range of estimates for temporal acuity of human ear perception. The size of each sub-segment was set to 30 milliseconds, where the FFT resolution was 2048 as the next-larger power-of-two number (NFFT). The harmonic peaks were estimated by searching for an FFT point of the local maximum amplitude in linear scale, which is close to a FFT point with a integer multiple of the fundamental frequency. Thus, the jth harmonic peak of the kth instrument in each frame can be

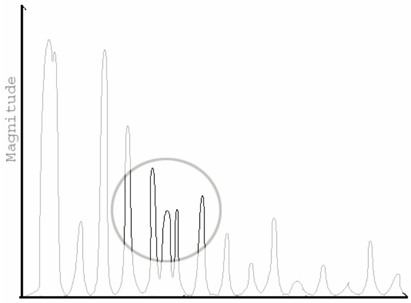

Figure 3. Harmonics of two different fundamental frequency are illustrated with a highlighted region where the clustering procedure shall be taken.

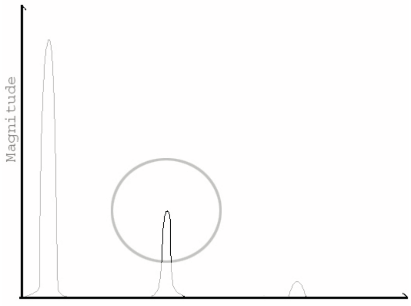

Figure 4. One set of separated harmonics with homogenous fundamental frequency.

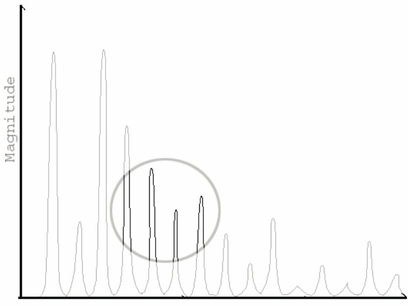

Figure 5. The other set of Separated harmonics with homogenous fundamental frequency.

computed in formula (10).

$$h_j^k = \max(\chi(n)), \quad n \in \left[\left\lfloor \frac{(j)l_k 2^{\lceil \log_2 w \rceil}}{S_r} \right\rfloor, \left\lceil \frac{(j)l_k 2^{\lceil \log_2 w \rceil}}{S_r} \right\rceil \right] \tag{1}$$

where l_k is the fundamental frequency of the kth instrument, $\chi(n)$ is the FFT coefficient. Here we limit the range of search for a local peak to avoid picking up a nearby peak of another sound source.

Consequently, k dominant instruments will result in k sets of harmonic peaks. Then, we put the resultant sets of harmonic peaks together to form a sequence of peaks H_j^p with a merge sort in an ascending order by the frequency.

In this sequence of peaks, for each pair of neighbor peaks, there are three possible situations: the two immediate peak neighbors are from the same sound source; the two immediate peak neighbors are from two different sound sources; part of one of the peak and the other peak are from the same sound source. The third case is due to two overlapping peaks, where the frequency is the multiplication of the fundamental frequencies of two different sound sources. In this scenario, the system first partition the energy between the two sound sources according to the ratio of the previous harmonic peaks of those two sound sources. Therefore, only the heterogenous peak will be processed.

The clustering algorithm has been used for separation of energy between two immediate heterogenous neighbor peaks (noted as H_j^p and H_{j+1}^p), which contains integer multiples of different fundamental

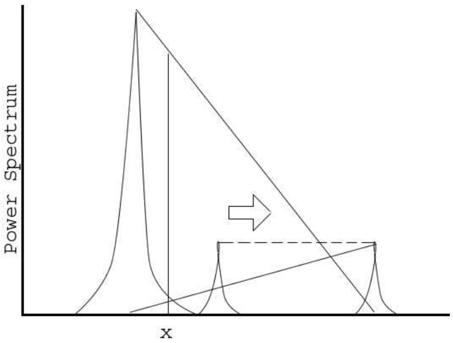

Figure 6. The illustration shows two peaks from different sound sources. The amplitude of the peak on the left side is bigger than that of the right one. The right peak is scaled to a point on its right side according to the amplitude ratio of those two peaks.

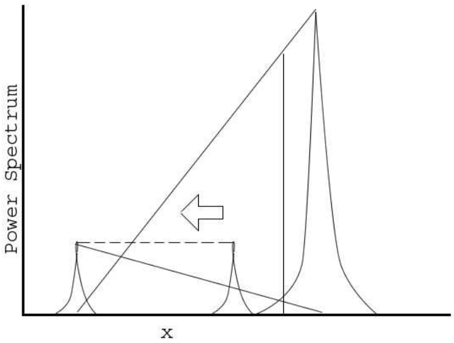

Figure 7. The illustration shows two peaks from different sound sources. The amplitude of the peak on the left side is smaller than that of the right one. The left peak is scaled to a point on its left side according to the amplitude ratio of those two peaks.

frequency. Considering the wide range of the magnitude of harmonic peaks, we applied a coefficient α to linearly scale each pair of immediate neighbor harmonic peaks to a virtual position along the frequency axis by a ratio of the magnitude values of the two harmonic peaks. Then the magnitude of each point between the two peaks was proportionally computed in each peak. The amplitude of the ith FFT point in the pth sub-segment can be represented as:

$$\chi_p(i) = \begin{cases} A_j^p \frac{\left[f(H_{j+1}^p)-i\right]\alpha}{\left[f(H_{j+1}^p)-i\right]\alpha+i-f(H_j^p)} + A_{j+1}^p \frac{i-f(H_j^p)}{\left[f(H_{j+1}^p)-i\right]\alpha+i-f(H_j^p)}, & A_j^p > A_{j+1}^p \quad (2) \\ A_j^p \frac{f(H_{j+1}^p)-i}{f(H_{j+1}^p)-i+\left[i-f(H_j^p)\right]\alpha} + A_{j+1}^p \frac{\left[i-f(H_j^p)\right]\alpha}{f(H_{j+1}^p)-i+\left[i-f(H_j^p)\right]\alpha}, & A_j^p < A_{j+1}^p \end{cases} \quad (3)$$

$$\alpha = \begin{cases} \sqrt{A_j^p / A_{j+1}^p}, & A_j^p > A_{j+1}^p \\ \sqrt{A_{j+1}^p / A_j^p}, & A_j^p < A_{j+1}^p \end{cases} \quad (4)$$

$$f(H_j^p) = \frac{H_j^p 2^{\lceil \log_2 w \rceil}}{S_r}, \quad j \in \left[1, \varepsilon / f_j\right], \quad (5)$$

where H_j^p is the frequency of the jth harmonic peak, $f(H_j^p)$ is the FFT point of the jth harmonic peak, A_j^p is the magnitude of the jth harmonic peak, and ε is an expected maximum frequency that western orchestral instruments can produce. For fast computation, a threshold for the magnitude of each FFT point has been applied, where only points with significant energy had been computed by the above formulas.

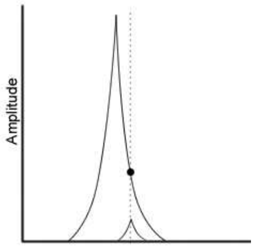

Figure 8. The illustration shows two estimated peaks from different sound sources. The real energy of one of them is insignificant. Clustering should not be applied here.

It is possible that an estimated harmonic peak may be assigned with energy that belongs to another sound source, which is close to it. See figure. This may seriously affect the clustering result. To overcome this problem, we defined that H_j^p is a peak only if $\sum_j^N H_j^p \gg \sum_j^N H_j^{p+1}$ to ensure that formula (2) has been properly implemented, where σ is a range of FFT points. Also, based on the fact that one sound source may be longer than others in a polyphonic sound, we assume that a musical instrument is not dominant only when its total harmonic energy is significantly smaller than the average of the total harmonic energy of all sound sources.

After clustering the energy, each FFT point in the analysis window has been assigned k coefficients Ψ_i^k for the dominant instruments accordingly.

$$\phi = \Psi Im\{A_i^k(j)\}, \quad \varphi = \Psi Re\{A_i^k(j)\}, \tag{6}$$

$$s^k(n) = \sum_{j=0}^{N-1} \chi_j(\varphi, \phi) e^{2ij\pi n/N}, n = 0, ..., N-1 \tag{7}$$

where $s^k(n)$ is the nth sample data of the kth instrument in each frame.

3. Feature Extraction

We consider the transient duration as the time to reach the quasi-steady state of fundamental frequency. Thus we only apply it to the harmonic descriptors, since in this duration the sound contains more timbre information than pitch information of the note, which is highly relevant to the fundamental frequency. The fundamental frequency is estimated by first computing the local cross-correlation function of the sound object, and then computing mean time to reach its maximum within each frame, and finally choosing the most frequently appearing resultant frequency in the quasi-steady status. In each frame i, the fundamental frequency is calculated in this form:

$$f(i) = \frac{S_r}{K_i/n_i} \tag{8}$$

where S_r is the sample frequency, n is the total number of $r(i,k)$'s local valleys across zero, where $k \in [1, K_i]$. See formula (9). K_i is estimated by k as the maximum fundamental period by the following formula, where $r(i,k)$ reaches its maximum value. ω is the maximum fundamental period expected.

$$r(i,k) = \sum_{j=1}^{m+n-1} s(j)s(j-k) \Bigg/ \sqrt{\sum_{j=m}^{m+n-1} s(j-k)^2 \sum_{j=m}^{m+n-1} s(j)^2}, \quad k \in [1, S_r \times \omega] \tag{9}$$

We observed that during the transients, the instantaneous fundamental frequencies are unstable, and usually very different from the one in the quasi-steady state, see fig. 3. Thus we choose to apply fixed frame size and hop size to all the sound objects instead of fundamental frequencies dependable sizes. In our project, frame size is set to 30 milliseconds, and hop size is set to 10 milliseconds, which is within the range of estimates for temporal acuity of human ear.

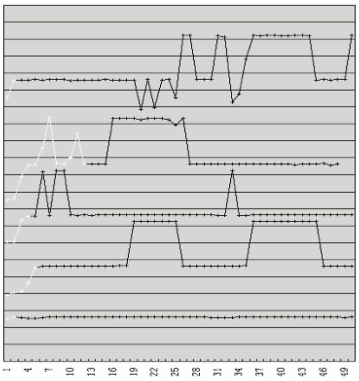

Figure 9. Fundamental frequency trajectories of sounds in Middle-C played by (bottom to top) alto flute (with vibrato), cello (bowed with vibrato), double bass (bowed), double bass (pizzicato), and flute (flutter) are displayed within first 50 frames. Highlightened line stands for the transient duration for each instrument. We observed that the transient duration of the Woodwind family is significantly shorter than that of the String family.

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 are 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, we applied three different empirical threshold values of time duration by the total length of each music object. For objects less than 30 frames, the threshold was set to three, which was a 30ms 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 ms and was more than 5% of its total length; for objects longer than 100 frames, the threshold was set to eight, which was 100ms. Also, we consider that the pitch is temporally stable, when the variation of an instantaneous fundamental frequency f_0^i of the ith frame, is less than $\theta\%$ the instantaneous fundamental frequency f_0^{i-1} of its previous frame, or less than $\theta\%$ octave ($2 * f_0^{i-1}$ or $f_0^{i-1}/2$) of its previous frame.

In this research, our time-frequency features are based on the MPEG-7 [11]low-level descriptors. They are as follows:

The Audio Power Spectrum Centroid describes the center-of-gravity of a log-frequency power spectrum in formula (10).

$$C(i) = \sum_n \log_2 \big(f(n)/1000\big) P^i(n) \Big/ P^i(n), \tag{10}$$

$$n = b, ..., \frac{NFFT}{2}, \quad b = \lfloor \frac{62.5NFFT}{S_r} \rfloor, \tag{11}$$

$$P^i(b) = \sum_{k=0}^{b} P^i(k), \quad f(b) = 31.25, \tag{12}$$

$$f(n) = (n)\frac{S_r}{NFFT}, \quad n = b, ..., B, \quad B = \frac{NFFT}{2} - b, \tag{13}$$

$$\bar{C} = \sum_{i=k}^{K} C_i \Big/ (K - k), \tag{14}$$

where $k = 0$, K is the last estimated transient frame for an average Power Spectrum Centroid in the transient state; $k = K$, K is the jth frame that the pitch stays stable within a threshold θ for the instantaneous fundamental frequency, and j is bigger than a time duration threshold.

The Audio Power Spectrum Spread describes the Root-Mean-Square deviation of the log-frequency power spectrum to its centroid in formula (15).

$$S(i) = \sqrt{\sum_{n=b}^{B} \Big(\big(\log_2(f(n)/1000\big) - C(i)^2 P^i(n) \Big) \Big/ \sum_{n=b}^{B} P^i(n)}, \tag{15}$$

$$\bar{S} = \sum_{i=k}^{K} S_i \Big/ (K - k), \tag{16}$$

where $C(i)$ is the instantaneous Power Spectrum Centroid in formula (10) in the ith frame.

The Harmonic Spectral Centroid describes a weighted mean of the amplitude in linear scale of the harmonic peaks of the spectrum. Estimation of harmonic peaks has been discussed previously in Section 2. The instantaneous Harmonic Spectral Centroid is calculated in formula (17).

$$HSC(i,s) = \frac{\sum_{j=1}^{N} h_j^s A_j^s}{\sum_{j=1}^{N} A_j^s}, \tag{17}$$

$$\overline{HSC(s)} = \sum_{i=k}^{K} HSC(i,s)/(K - k), \tag{18}$$

where h_j^k is the frequency of the jth harmonic peak for the sth instrument.

The Harmonic Spectral Deviation describes the spectral deviation of the amplitude in logarithmic scale of the harmonic peak from a spectral envelope. The instantaneous Harmonic Spectral Deviation in formula (19).

$$HSD(i,s) = \frac{\sum_{j=1}^{N} \left| \log_{10} A_j^s - \log_{10} SE \right|}{\sum_{j=1}^{N} \log_{10} A_j^s}, \tag{19}$$

$$SE = \frac{\sum_{i=-1}^{1} A_{j+i}^s}{3}, \quad j \in [-1, K], \tag{20}$$

$$\overline{HSD(s)} = \sum_{i=k}^{K} HSD(i,s)/(K-k), \tag{21}$$

where $A_j^s = 0$, if$j = 0$ or $j = K$.

The Instantaneous Harmonic Spectral Spread describes the linearly-scaled amplitude weighted standard deviation of the harmonic peaks from the harmonic centroid, normalized by its instantaneous Harmonic Spectral Centroid in formula (24).

$$HSS(i,s) = \frac{1}{HSC(i,s)} \sqrt{\frac{\sum_{j=1}^{N} (A_j^s)^2 \big(h_j^s - HSC(i,s)\big)^2}{\sum_{j=1}^{N} (A_j^s)^2}}, \tag{22}$$

$$\overline{HSS(s)} = \sum_{i=k}^{K} HSS(i,s)/(K-k), \tag{23}$$

The Instantaneous Harmonic Spectral Variation describes a normalized correlation between the linearly-scaled amplitude of the harmonic peaks of a pair of immediate frame neighbors in formula (4).

$$HSV(i,s) = 1 - \frac{\sum_{j=1}^{N} (A_j^{p-1}) A_j^p}{\sqrt{\sum_{j=1}^{N} (A_j^{p-1})^2} \sqrt{\sum_{j=1}^{N} (A_j^p)^2}}, \tag{24}$$

$$\overline{HSV(s)} = \sum_{i=k}^{K} HSV(i,s)/(K-k), \tag{25}$$

where A_j^p represents amplitude in the linear scale of the jth harmonic peak of the sth musical instrument, in the pth frame.

The Spectral Centroid describes the weighted average amplitude of the frequency bins in the power spectrum for each frame in formula (28).

$$SC(i) = \frac{\sum_{j=1}^{N} A_j f(j)}{\sum_{j=1}^{N} A_j}, \tag{26}$$

$$\overline{SC} = \sum_{i=k}^{K} SC(i)/(K-k), \tag{27}$$

where N is the total number of bins of the power spectrum.

The Temporal Centroid describes the time averaged over the signal envelope in formula (6).

$$TC = \frac{\sum_{n=1}^{N} nSE_{nv}(n)/S_r}{\sum_{n=1}^{N} SE_{nv}(n)}, \tag{28}$$

$$SE_{nv}(n) = \sum_{i}^{K} [s(n,i)]^2/K, \tag{29}$$

where N is the total number of analysis windows, and K is the size of the window.

The Descriptors for the whole segment include Mean Spectrum Centroid, Mean Spectrum Spread, Log Attack Time, Mean Spectral Centroid, Temporal Centroid, Mean Spectral Centroid, Transient Duration; descriptors for transient state, quasi-steady state and the whole segment include TS Harmonic Spectral Centroid, TS Harmonic Spectral Deviation, TS Harmonic Spectral, TS Harmonic Spectral Spread, TS Spectral Centroid, QS Harmonic Spectral Centroid, QS Harmonic Spectral Deviation, QS Harmonic Spectral Variation, QS Harmonic Spectral Spread, QS Harmonic Spectral, QA Harmonic Spectral Centroid, QA Harmonic Spectral Deviation, QA Harmonic Spectral Variation, QA Harmonic Spectral Spread, where TS represents Transient; QS represents the Quasi-Steady state; QA represents the whole segment, which includes both the transient state and the Quasi-Steady state. Each descriptor has a single value.

4. Classifiers

The classifiers, applied in the investigations on musical instrument recognition, represent practically all known methods. In our research, we began with, but not limited to, four classifiers (Bayesian Networks, Logistic Regression Model, Decision Tree J-48 and Locally Weighted Learning) upon a group of samples from our music sound object database to explore the effectiveness of our new descriptors.

Bayesian Networks is a widely used statistical approach, which represent 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:

$$p(x_1, ..., x_N) = \prod_{i=1}^{N} p(x_i \mid x_{\pi i}) \tag{30}$$

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

Logistic regression model is a popular statistical approach of analyzing multinomial response variables, since it does not assume normally distributed conditional attributes which can be continuous, discrete, dichotomous or a mix of any of these; it can handle nonlinear relationships between the decision attribute and the conditional attributes. It has been widely used to correctly predict the category of outcome for new instances by maximum likelihood estimation using the most economical model. A generalized logit model has a form:

$$\log\left(\frac{P_r(Y=i\mid x)}{P_r(Y=k+1\mid x)}\right)=\alpha_i+\beta_i x_i,\quad i=1,...k \tag{31}$$

where the $k+1$ possible responses have no natural ordering and the α_1, ... , α_k are k intercept parameters, and the β_1, ... , β_k are k vectors of parameters. For details, see [8].

Locally weighted Regression is a well-known lazy learning algorithm for pattern recognition. It votes on the prediction based on a set of nearest neighbors (instances) of the new instance, where relevance is measured by a distance function. The local model consists of a structural and a parametric identification, which involve parameter optimization and selection. For details see [6].

Decision Tree-J48 is a supervised classification algorithm, which has been extensively used in machine learning and pattern recognition. See [20], [25]. 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.

5. System Overview

Our system is composed by six modules: a STFT convertor with hamming window, a harmonic peak estimator, an energy clustering engine with matrix factorization, an inverse FFT engine with an input of a sets of scaled complex numbers and an output of the sample data, a feature extractor and four classifiers.

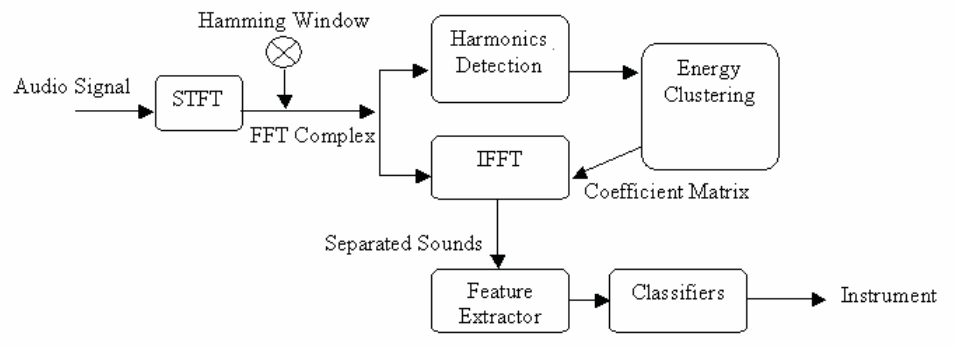

Figure 10. System architecture with input of a polyphonic sound object and output of the instrument / family estimation.

A STFT convertor: Divide a digital audio object into a sequence of frames, apply STFT transform to the mixed sample data of integers from time domain to the frequency domain with a hamming window, and output NFFT discrete points.

A harmonic peak estimator: Compute harmonic peaks by searching for local peaks around the integer multiples of the fundamental frequencies, compare the total energy of each musical sound source: ignore harmonics from a sound source if their total energy is insignificant, compare to others.

An energy clustering engine: Merge-sort sets of harmonic peaks into one sequence by its frequency value, proportionally calculate the energy of each FFT point between each pair of heterogeneous harmonic peaks and output a matrix of coefficients.

An inverse FFT engine: Multiply the complex numbers of the FFT points by the coefficient matrix and transform them back to the time domain in a form of separate audio files.

A feature extractor: Detect the transient-steady border, retrieve MPEG-7-based descriptors, and average the feature values in different states.

Four different classifiers: Train the classifiers by the monophonic sound in out database, and classify the separated sounds.

6. Experimental Design

We used a database of 432 music recording sound objects with a full coverage of the pitch range of 7 instruments (violin, viola, cello, bass flute, flute, clarinet and alto flute) for two instrument families (string and woodwind), from McGill University Master Samples CD Collection, which has been widely used for research on musical instrument recognition all over the world. All sounds have a sample rate of 44,100Hz with mono channel. Each type of instrument may have up to two different articulations of samples. From each group of sound which was grouped by the instrument type, 4 sounds of different pitches were randomly chosen for mixing. Totally 24 different monophonic sounds were mixed in pairs by summation for the purpose of testing our sound separation algorithm. All the 12 polyphonic sounds were produced by pair of sounds from different instrument families. Therefore, there were totally 24 separated sounds in the testing set, while all the 432 music recording sound objects were used in the training set. In this research, we only focused on classification of the instrument family, see (6). We used WEKA for all classifications.

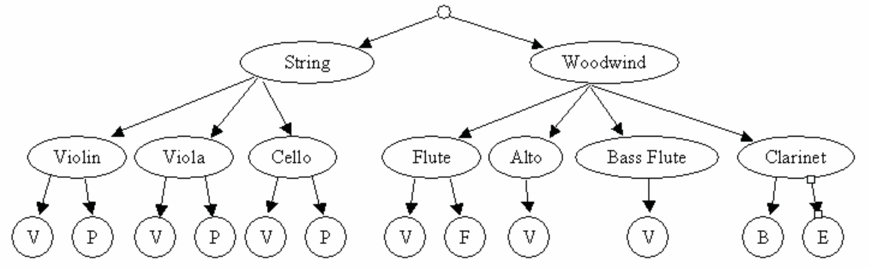

Figure 11. Sound objects in our experiment are of 7 instruments (from top to bottom: bass flute, piccolo, alto flute, flute, cello, viola, violin). B represents b-flat; E represents e-flat; P represents pizzicato; V represents vibrato; F represents flutter.

$Instrument$	$Type$	$Note$
Clarinet	B-Flat	3A
Violin	bowed vibrato	4C

Table 1. An example of a sound mix.

7. Results

Our classification results are displayed in Table 7. In each row, the classification correctness of the each set of feature vectors over different classification algorithms is presented. ”TS” is a set of feature vectors with harmonic features from the transient state TS Harmonic Spectral Centroid, TS Harmonic Spectral Deviation, TS Harmonic Spectral Variation, TS Harmonic Spectral Spread, Log Attack Time, Mean Spectral Centroid, Temporal Centroid. ”QS” is a set feature vectors with harmonic features from the quasi-steady state QS Harmonic Spectral Centroid, QS Harmonic Spectral Deviation, QS Harmonic Spectral Variation, QS Harmonic Spectral Spread, Log Attack Time, Mean Spectral Centroid, Temporal Centroid. ”GA” is a set of feature vectors with harmonic features of the whole segment GA Harmonic Spectral Centroid, GA Harmonic Spectral Deviation, GA Harmonic Spectral Variation, GA Harmonic Spectral Spread, Log Attack Time, Mean Spectral Centroid, Temporal Centroid.

	BN	LM	LWL	$Tree48$
TS+QS	58.33 %	45.83 %	66.67 %	58.33 %
TS+GA	58.33 %	54.17 %	66.67 %	70.83 %
QS+GA	75.00 %	58.33 %	62.50 %	70.83 %
TS+QS+GA	70.83 %	58.33 %	66.67 %	70.83 %
TS	62.50 %	50.00 %	66.67 %	54.17 %
QS	66.67 %	58.33 %	62.50 %	66.67 %
GA	66.67 %	54.17 %	62.50 %	70.83 %
Ave.	65.48 %	54.17 %	64.88 %	60.07 %

Table 2. Sound separation of sound mixes from different instrument families. Classification: violin against clarinet. BN stands for Bayesian Networks. LM stands for Logistic Regression Model. LWL stands for Local weighted learning.

8. Conclusion and Future Work

We observed that the harmonic features in transient state of music sound objects behave significantly different from those in the quasi-steady state. Since we consider the transient duration as the time to reach the stable state of the pitch of a note, the transient state contains instrument signature information with less pitch information. Therefore, based on the fact that human perception relies on the transient part, data in the transient state is highly correlated to the timbre properties and less relevant to the funda-

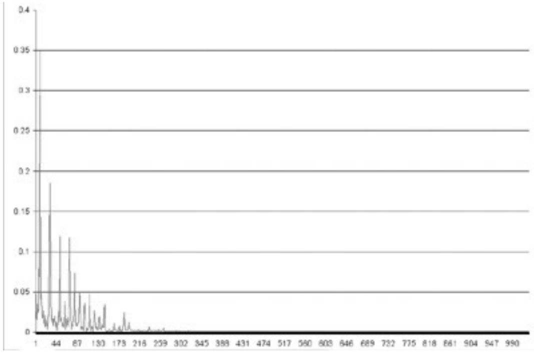

Figure 12. Shows the power spectrum in quasi-steady state (fifth frame) of a successfully classified 3A clarinet sound, which was separated from the example sound mix.

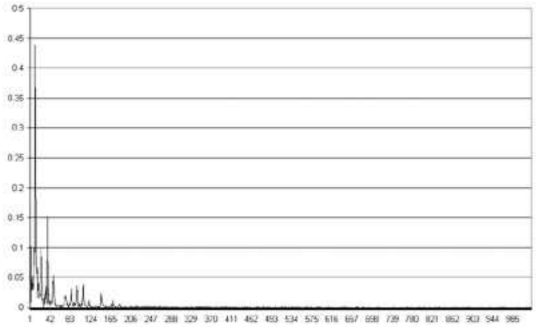

Figure 13. Shows the power spectrum in quasi-steady state (fifth frame) of an unsuccessfully classified 4C violin sound, which was separated from the example sound mix.

mental frequency of the pitch. Thus, the clustering algorithm fits better in the quasi-steady period than in the transient period, where the harmonics were computed based on the instantaneous fundamental frequencies. However, the group of features in the quasi-steady period alone is not adequate for sound classification, as shown in our results. Together with the segment-wise features, the group of features has better accuracy.

For classification of the instrument families, the Bayesian Networks classifier has the best overall performance while the Logistic Regression Model algorithm has the poorest performance among all the classifiers that we were used in the experiments.

The proposed research is still in its early stages. More new features, which represent acoustical behaviors, and more music objects, which are from more instruments families in different articulations, shall be investigated. Also, future research shall explore the efficiency of more classifiers, too.

9. Acknowledgment

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

References

[1] A. Wieczorkowska, J. Wroblewski, P. S., Slezak, D.: Application of Temporal Descriptors to Musical Instrument Sound, *Journal of Intelligent Information Systems*, **21**(1), 2003, 71–93.

[2] Back, A. D., Weigend, A. S.: A first Application of Independent Component Analysis to Extracting Structure from Stock Returns, *International Journal onNeural Systems*, **8**(4), 1998, 474–484.

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

[4] Bregman, A. S.: *Auditory scene analysis, the perceptual organization of sound*, vol. 3, MIT Press, 1990.

[5] Brown, J. C.: Musical instrument identification using pattern recognition with cepstral coefficients as features, *Journal of Acousitcal society of America*, **105**(3), 1999, 1933–1941.

[6] C. G. Atkeson, A. W. M., Schaal, S.: Locally Weighted Learning for Control, *Artificial Intelligence Review*, **11**(1-5), Feb 1997, 75–13.

[7] Cadoz, C.: Timbre et causalite, Apr 1985, Seminar on Timbre, Institut de Recherche et Coordination Acoustique/Musique, Paris, France.

[8] le Cessie, S., van Houwelingen, J.: Ridge Estimators in Logistic Regression, *Applied Statistics*, **41**(1), 1992, 191–201.

[9] Eronen, A., Klapuri, A.: Musical instrument recognition using cepstral coefficients and temporal, the IEEE International Conference on Acoustics, Speech and Signal Processing. Istanbul, Turkey, June.

[10] Fujinaga, I., McMillan, K.: Realtie Recognition of Orchestral Instruments, International Computer Music Conference, 2000.

[11] Group, M. P. E.: Information Technology - Multimedia Content Description Interface, Part 4: Audio, International Organizaiton For Standardization. ISO/IEC JTC 1/SC 29/WG 11.

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

[13] Jensen, K., Arnspang, J.: Binary decision tree classification of musical sounds, the 1999 International Computer Music Conference, Beijing, China, Oct.

[14] K. Livescu, J. G., Bilmes, J.: Hidden Feature Models for Speech Recognition Using Dynamic Bayesian Networks, Eurospeech, Geneva, Switzerland, Sep 2003.

[15] Kaminskyj, I., Materka, A.: Automatic source identification of monophonic musical instrument sounds, 1, the IEEE International Conference On Neural Networks, 1995.

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

[17] Martin, K. D.: *Sound-Source Recognition: A Theory and Computational Model*, Ph.D. Thesis, MIT, Cambridge, MA,.

[18] Martin, K. D., Kim, Y. E.: Musical instrument identification: A pattern-recognition approach, the 136th meeting of the Acoustical Society of America, Norfolk, VA, 1998, 2pMU9.

[19] Pollard, H., Jansson, E.: A Tristimulus Method for the Specification of Musical Timbre, *Acustica*, **51**(5), 1982, 162–171.

[20] Quinlan, J. R.: C4.5: Programs for Machine Learning, 1993, Talk at San Mateo, CA: Morgan Kaufmann.

[21] Ristaniemi, T., Joutsensalo, J.: On the Performance of Blind Source Separation in CDMA Downlink, 34, Workshop on Independent component Analysis and Signal Separation(ICA'99), 1999.

[22] Smaragdis, P.: *Redundancy Reduction for Computational Audition, A Unifying Approach*, Ph.D. Thesis, Media Arts and Sciences Department, MIT.

[23] Virtanen, T., Klapuri, A.: Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, 2001.

[24] Wieczorkowska, A.: *The Recognition Efficiency of Musical Instrument Sounds Depending on Parameterization and Type of a Classifier*, Ph.D. Thesis, Technical University of GDansk, Poland.

[25] Wieczorkowska, A.: Classification of musical instrument sounds using decision trees, in the 8th International Symposium on Sound Engineering and Mastering, ISSEM'99, 1999.

[26] Zweig, G.: *Speech Recognition with Dynamic Bayesian Networks*, Ph.D. Thesis, Univ. of California, Berkeley, California, 1998.