

# Harmonic Blind Sound Source Isolation Enhanced by Spectrum Clustering

Xin Zhang  
University of North Carolina  
Dept. of Computer Science  
Charlotte, NC 28223  
704-687-8546  
xinzhang@uncc.edu

Wenxin Jiang  
University of North Carolina  
Dept. of Computer Science  
Charlotte, NC 28223  
704-687-8546  
wjiang3@uncc.edu

Zbigniew W. Ras  
University of North Carolina  
Dept. of Computer Science  
Charlotte, NC 28223  
704-687-8546  
ras@uncc.edu

## ABSTRACT

*Automatic indexing of music by instruments and their types is a challenging problem, especially when multiple instruments are playing at the same time. We have built a database containing more than one million of music instrument sounds, each described by a large number of features including standard MPEG7 audio descriptors, features for speech recognition, and many new audio features developed by our team. Our previous research results show that all these features only lead to classifiers which successfully identify music instruments in monophonic music (only one instrument playing at a time). Their confidence for polyphonic music is much lower. This brought the need for blind sound source separation algorithms. In this paper, we present a new spectrum clustering enhanced method which improves the estimation of fundamental frequency as well as the balance of the categorization tree of training datasets, and therefore enhances the precision of automatic indexing. The system recursively detects the pitch of the predominant sound source, then calculates the features based on the estimated pitch, then predicts the most similar spectrum by the corresponding classification tree, and finally subtracts the estimated predominant spectrum from the database until silence is detected.*

## 1. INTRODUCTION

In recent years, rapid advances in digital music creation, collection, and storage technologies have enabled organizations to accumulate vast amounts of musical audio data. The booming of multimedia resources on the Internet has brought a tremendous need to provide new, more advanced tools for the ability to query and process vast quantities of musical data, since searching through multimedia data is a highly nontrivial task requiring content based indexing of the data, which are not easy to describe with mere symbols. Lots of multimedia-resources provide data, 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. Timbre recognition, one of the main subtasks of Music Information Retrieval, has proven to be extremely challenging especially in multi-timbral sounds, where different instruments are playing at the same time in the same channel of a digital music recording.

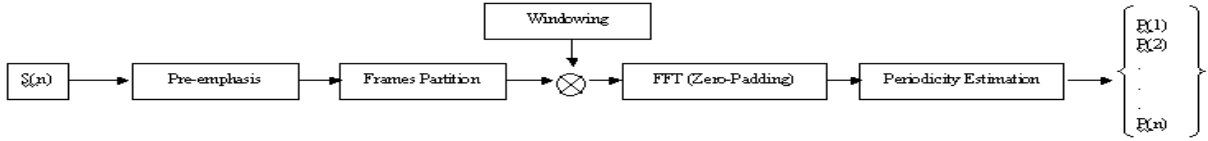
Intensive investigations have been made by a number of researchers to develop computational features to describe the characteristics of monophonic sound pieces where there is only one sound source per sound object per channel. Recently, the Moving Picture Expert Group (MPEG) has published the MPEG7 standard of a set of acoustical features based on the latest research in this area. However, most of these features failed to describe sufficient information to distinguish timbers in multi-timbre sounds, where multiple sound sources are active at the same time. This brought the need for Blind Sound Separation to pre-process multi-timbre sounds into monophonic sounds before feature extraction. Human hearing perception system can focus on a few sound sources in a multi-sounds environment where different musical instruments are playing at the same time. However, it is a very challenging task for a computer to recognize pre-dominant musical sound sources in sound mixtures, which is also called a Cocktail Party Problem [9]. In the next subsection, we give a review on Multi-pitch estimation and Blind Signal Separation, since the research presented in this paper has implications for studies in blind harmonic sound separation and pre-dominant fundamental frequency estimation.

### 1.1. Pitch Estimation in Frequency Domain

Pitch detection has been extensively explored by lots of audio signal processing researchers ([22], [32], [2], and [10]). Pitch detection techniques have been widely used in music transcription and music file annotation. Numerous methods of pitch detection have been developed and

explored, which can be categorized by the functional domain into three different types: time, frequency, and time-frequency. The authors focus on reviewing the most promising type of the fundamental frequency estimation algorithms, which leads to predominant-pitch detection: pitch estimation in the frequency domain. Most known and well-established algorithms in other domains, such as autocorrelation [2] and Average Magnitude Difference Function [10] in the time domain, which have been successfully applied in mono-signal processing, fail to detect the fundamental frequencies of sound mixtures in multi-timbre sounds.

Many interesting methods have been explored by a number of researchers to detect fundamental frequency in the frequency domain ([5], [16], [3], [2] and [15]). The diagram of a common frequency domain pitch detector is shown in Figure 1. One approach is to use a group of hypothetical fundamental frequencies for a comb function [5], where the fundamental frequency is estimated by a hypothetical fundamental frequency that maximized the value of a sum of products of the comb function and its corresponding power in the STFT spectrum:



**Figure 1. Common frequency domain pitch detector diagram.**

$$C(m, f_h) = \begin{cases} 1, m = kf_h, k \in [1, N] \\ 0, m \neq kf_h \end{cases} \quad (1.)$$

$$A_c(f_h) = \sum_{k=1}^{\frac{N}{2f_h}} X(kf_h) \times C(kf_h, f_h)$$

where  $kf_k$  is the highest integer multiple of the  $k$ th candidate frequency smaller than half the sampling rate  $N$ , and  $X$  is the power of the spectrum.

Beauchamp et al. extended this method by replacing the comb function with a two-way mismatch function [2]:

$$e_1 = \sum_i^{\frac{N}{2f_h}} \min_x |if_h - f_x| \rho(f_x, A_x)$$

$$e_2 = \sum_x^M \min_i |if_h - f_x| \rho(f_x, A_x) \quad (2.)$$

$$E = w_1 e_1 + w_2 e_2$$

where  $N$  is the sampling rate,  $w_1$  and  $w_2$  are empirical coefficients. The drawback of this type of strategy is that the selection of a group of hypothetical fundamental frequencies is critical to their system performance and efficiency.

Another approach is based on the Schroeder's histogram method, which uses the maximum value in the Schroeder's histogram of the integer multiples of each peak frequency to estimate the fundamental frequency [28]. Hess extended this approach by applying a compressed spectrum to the histogram [16]. Edgar et al. improved this algorithm with a maximum likelihood function by taking the distance between the real peak and

the integer multiple of a candidate fundamental frequency and the priority of the frequency order into account [3].

$$f_0 = \max \left\{ \sum_{i=1}^k (C \log A_i) \left( C_e \frac{d_i^2 + f_i}{D} \right) \right\} \quad (3.)$$

The above review is not a complete for all the fundamental frequency estimation. It focuses on the pitch detection by the frequency components in the power spectrum.

## 1.2. Sound Source Estimation

Blind Signal Separation is a very general problem in lots of areas besides musical sound timbre recognition: neural computation, finance, brain signal processing, general biomedical signal processing, speech enhancement, etc. Numerous intersecting techniques have been investigated in this area, which can be categorized into, but not limited to the following types: Filtering Techniques ([30], [1], [4]), Independent Component Analysis (ICA) ([17], [11], [7]), the Degenerate Unmixing Estimation Technique (DUET) [20], Factorial Hidden Markov Models (HMM) [26], Singular Value Decomposition (Spectrum Basis Functions in MPEG7 [18]) and Harmonic Sources Separation Algorithms ([12], [25] and [29]). Filtering Techniques, ICA and DUET require different sound sources to be stored separately in multiple channels. Most often, HMM works well for sound sources separation, where fundamental frequency range is small and the variation is subtle. However, unfortunately, western orchestral musical instruments can produce a wide range of fundamental frequencies with dynamic variations.

Spectral decomposition is used to efficiently decompose the spectrum into several independent subspaces [8] with smaller number of states for HMM. Commonly, Harmonic Sources Separation Algorithms have been used to estimate sound sources by detecting their harmonic peaks, decoding spectrum into several streams and re-synthesizing them separately. This type of methods relies on multi-pitch detection techniques and iterative Sinusoidal Modeling (SM) [12]. For the purpose of interpolating the breaks in the sinusoidal component trajectories, numerous mathematical models have been explored: linear models [31], and non-linear models such as high degree interpolation polynomials with cubic spine approximation model [12], etc. However, it is very difficult to develop an accurate sinusoidal component model to describe the characteristics of musical sound patterns for all the western orchestral instruments. In this research, we focus on separating harmonic sound signal mixtures in a single channel by isolating and matching the pre-dominant harmonic features with connection to a feature database. In terms of applying harmonic signature information to distinguish timbre, our sound separation method is similar to the SM approach. However, instead of using a model to describe an input signal, we applied decision tree to estimate the most similar harmonic signature in the feature database, and then identified its pre-stored spectrum. Given an unknown sound mixture, our sound separation system first identifies pre-dominant pitch among a set of harmonic candidate peaks by an enhanced maximum likelihood algorithm based on spectrum clustering technique, and then compares a sequence of the corresponding harmonic peaks with those in our feature database and estimates the unknown sound source by the classifiers, and then subtracts the matched sound from the unknown sound mixture, and repeats the same steps to the remaining signal. The following sections begin with an outline of our system, and then describe the details of algorithm in this research.

## 2. HARMONIC SOURCE ISOLATING SYSTEM

Our system consists of six modules: a quasi-steady state detector, a STFT converter with hamming window, a pre-dominant fundamental frequency estimator, a decision tree classifier with connection to a feature database of instantaneous spectral harmonic features, a temporal harmonic feature-based frame-matching device, and a spectrum subtraction engine with an audio database, as shown in Figure 2. First, the system computes the predominant pitch for each frame and identifies the beginning of the stable state of predominant pitch in an input digital musical file, where, for each frame having a

stable pitch, the system computes and outputs a vector of harmonic features to a decision tree classifier. Second, the classifier identifies the timbre and articulation of predominant sound source within the frame, which leads to a sample musical file in the audio database. Then the frame estimation device identifies the closest-matched frame in the estimated musical sample file by a pitch-related temporal feature. Finally, it subtracts the spectrum in the frame of the input file from that in the estimated frame of the estimated file. It repeats the steps until either a threshold is reached, which indicates no harmonic sound sources left, or the maximum loop is reached. The figure below shows an overview of the system. The quasi-steady state detector sorts out unstable frames with insignificant harmonic characters, by computing predominant fundamental frequency of each frame and outputting the beginning and the end cuts among the frames. The STFT converter divides a digital audio object into a sequence of frames, applies STFT transform to the mixed sample data of integers from time domain to frequency domain with a hamming window, and outputs NFFT discrete points.

The predominant fundamental frequency estimator identifies all the possible harmonic peaks and categorizes them into two clusters: one for predominant candidate; the other for none-predominant candidate. Then, for the predominant group, it computes the likelihood value for each peak, elects the frequency with the maximum likelihood value as the fundamental frequency, and stores its normalized correspondence harmonic sequence.

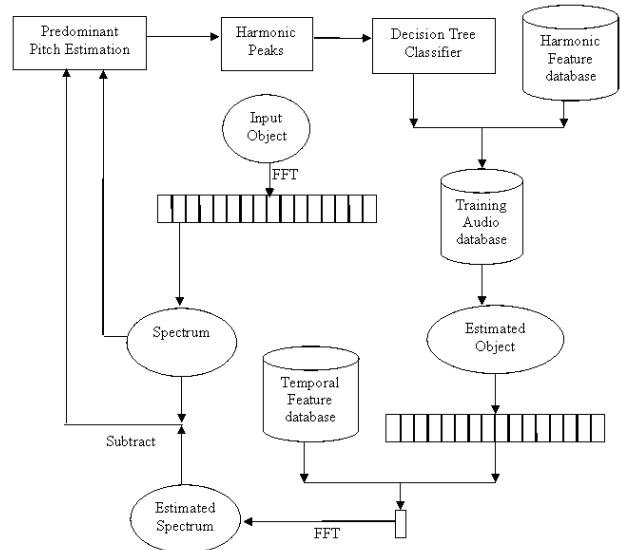


Figure 2. Sound separation system overview.

The decision tree classifiers compute an estimation of the sound source, with which the estimated pitch can associate the input file to a sample music object in the database.

The FFT subtraction device subtracts the detected sound source from the spectrum, computes the imaginary and real part of the FFT point by the power and phase information, performs IFFT for each frame, and outputs resultant remaining signals into a new audio data file.

### 2.1. Steady-Pitch State Estimation

This research investigates harmonic sequence information for the purpose of distinguishing the sound timbre, where energy is significantly distributed in mathematical relationships and fundamental frequency variation is relatively subtle. Also, by focusing on the steady frames, it efficiently shrinks down the size of the feature database for the purpose of pattern matching.

The beginning of the steady pitch state is at the first frame having an overall fundamental frequency variation in a desirable range among its continuous following  $N$  neighbor frames, where the total energy in the spectrum is bigger than a threshold in the case of silence ( $N$  was an empirical value,  $N=4$ ). Each frequency bin corresponds to a music note. The overall fundamental frequency is estimated by predominant fundamental frequency.

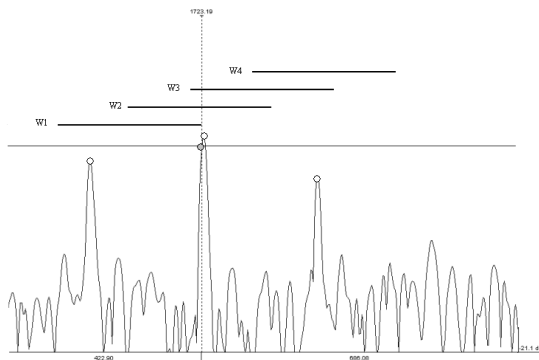
### 2.2. Predominant Music Pitch

In each steady-quasi frame, pre-dominant fundamental frequency is elected among a group of harmonic peaks by a maximum likelihood function. A peak is defined as a point having power value bigger than its immediate neighbor FFT points. Harmonic peaks are estimated by convolution window of mean amplitude greater than flexible threshold  $t$ .

$$P_i > t, t = C \cdot A_{max}, A_{max} > A_{silence} \quad (4.)$$

$$\chi_i^P > \chi_{i-1}, \chi_i^P > \chi_{i+1} \quad (5.)$$

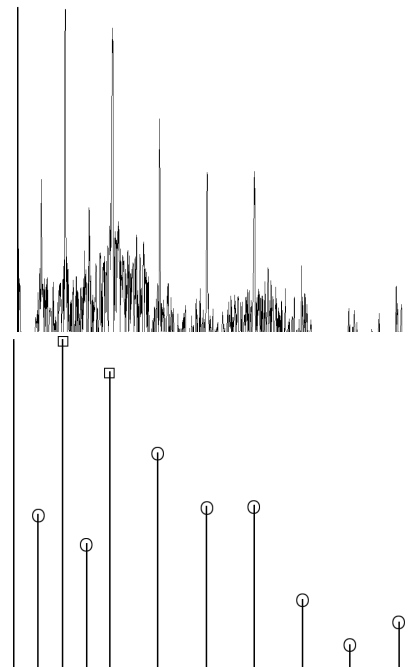
where  $C$  is a empirical coefficient,  $A_{max}$  is the largest amplitude in the spectrum,  $P_i$  is the  $i^{th}$  candidate peak,  $\chi_i$  is the power of  $P_i$ ,  $\chi_{i-1}$  is the power of the FFT point before  $P_i$ , and  $\chi_{i+1}$  is the power of the FFT point after  $P_i$ .



**Figure 3. Candidate peak identification.**

Figure 3 shows that four peak candidates are identified by four continuous windows, one peak per window. The peak marked with gray circle from window 1 (displayed as  $w1$  in the figure) is removed from the candidate list, since it doesn't meet the requirement in the above equation for power value: its power value is less than the neighbor FFT point to the right. Since our research interest is focused on finding the predominant harmonic sound source in each round, a single cut was needed to separate peak candidates into two clusters: one is the group of predominant harmonic peaks; the other is a group of none-predominant peaks.

The figure below shows that the candidate peaks in the 4<sup>th</sup> frame of a music sound played by piccolo in C# of 7<sup>th</sup> octave (theoretically associated with 2217.5Hz), where a rectangle mark at the top of a peak represents predominant, and a circle mark stands for the none-predominant. Therefore, the first candidate peak (1111.0Hz) was ruled out for predominant pitch calculation so that estimated pitch will be in the range between the second (2248.8Hz) and the fourth (4492.3Hz) candidate. K-means clustering algorithm was applied to grouping the most significant peak by pitch.



**Figure 4. Clusters generated based on magnitude.**

Each selected harmonic peak is then treated as a fundamental frequency candidate, where only harmonic peaks in higher ordinal position will be considered as its possible corresponding harmonic series. This algorithm favors those candidate peaks in lower ordinal positions

with higher priority for accumulative weights of magnitude in a mini-decibel scale. For each candidate peak, the weight is computed by the following equation:

$$W = \sum_i^{Sr/F_i} (30 + 10 \log_{10}(A'_i)) \quad (6.)$$

$$A'_i = \max\{A_k\}, k \in [i - c, i + c] \quad (7.)$$

$$c = \frac{F_k}{2^{\frac{11}{12}} \cdot df} \quad (8.)$$

where  $Sr$  is the sampling rate,  $F_i$  is the frequency of the  $i^{\text{th}}$  candidate peak,  $c$  is a range of the possible corresponding harmonic peak,  $df$  is the degree of freedom of FFT. Finally, the magnitude of each harmonic peak is normalized by the summation of those of the first  $N$  harmonic peaks.

### 2.3. Features and Classifiers

After the system detects the pre-dominant fundamental frequency, it computes a vector of normalized harmonic features  $F$ .

$$F_i = \frac{P_i}{P_0}, i = 1, \dots, N \quad (9.)$$

where  $P_0$  is the power of the predominant fundamental frequency,  $P_i$  is the power of the  $i^{\text{th}}$  harmonic partial of it in a frame,  $N=10$ .

Popular classifiers, which have been applied to automatic musical sound classification and speech recognition, include the  $k$ -Nearest Neighbors algorithm ([21], [14]), Naive Bayesian Classifiers ([24], [6]), Decision Trees ([19]), Discriminant Analysis ([36]), Higher Order Statistics, Artificial Neural Networks, Support Vector Machines, Rough Sets and Hidden Markov Models, etc. Hierarchical classification procedures together with the  $k$ -NN algorithm and other classifiers have been explored in different sound recognition systems ([23], [13]), 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. In the light that different timbre present harmonic patterns highly correlated to octaves, decision tree classifier [27] is used to identify the timbre and articulation (instrument playing method) in a frame to associate the frame to a similar musical file in the audio database, where harmonic features are automatically selected by different classifiers for different pitch as well as octave. In the feature database, the harmonic feature of each frame is treated as an observation of a music object,

while each music file is treated as a class of sound source containing instrument and articulation.

To minimize the distraction of pitch, the authors grouped the feature database by the estimated pitch and constructed a classifier for each group, where in each group all the training data records had the same estimated music note. For example, {A3, Flute, flutter, MUMS} is a class located in the database of pitch A3, where each frame in the steady state of the sound file "A3\_flute\_flutter\_M.au" is used as a training object. In the A3 database for the construction of this particular classifier, the pitch of every training object is A3, while the timbre and playing method of every training object may be different from each other. Intuitively, an accurate pitch estimation system may bring a balanced categorization schema; otherwise, a classifier of a particular pitch category with too many training objects incorrectly categorized may obstruct sound isolation by decreasing the efficiency of the classifier.

### 2.4. Frame Estimation

After the classifier identifies the most similar musical object in the database, it estimates the closest-matched frame in the estimated musical object for the current frame in the input file. The estimation is done by comparison of a temporal feature  $R$ :

$$R_k = \frac{P_0^k}{P_0^{k+m}}, k = 1, \dots, N \quad (10.)$$

where  $P_0^k$  is the power of the predominant fundamental frequency in the  $k^{\text{th}}$  frame of a music object,  $N$  is the total number of frames in stable pitch,  $m$  is a lag,  $m=5$ .

The frame having the smallest difference of  $D$  is chosen as the closest-matched one.

$$D_k = |R_k - R'_k|, k = 1, \dots, N \quad (11.)$$

where  $R_k$  is the ration of the frame in the input file,  $R'_k$  is that of the frame in the estimated music object.

### 2.5. Spectrum Subtraction

The harmonic sequence of estimated sound from the database is subtracted from the unknown sound by the real part. The imaginary part is then computed by the phase information of the input unknown sound.

### 2.6. Feature Database

Underlying the system is a large database of harmonic features, which was calculated from the sound objects originated from the McGill University Master Samples (MUMS) in form of AU format. Every musical sound object had multiple frame-based records to describe its

harmonic information within the spectrum. A last frame of a sound file, in which the total number of samples is less than the frame size, was truncated, due to the fact that it may not contain enough information to correctly describe the periodicity pattern, even though it may be in the steady-quasi state. These harmonic peak sequences were grouped by their music notes. Their columns include an audio file name, a frame number, and a peak identification number. The database covers the entire pitch range of all the harmonic music instruments in the western orchestra.

### 3. EXPERIMENTS

The system was implemented by Microsoft Visual Studio .NET, Microsoft SQL Server, FLTK, and OpenGL. The sampling rate is 44,100Hz, which is a common rate in musical compact disks. A frame size of 120 milliseconds was chosen with a hop size of 40 milliseconds. The next larger integer (NFFT) of the spectrum was 8192. In the training database of 3737 harmonic sound objects, there are totally 97 different music notes, where each musical note was played by a group of different musical instruments with different play methods. We used over 57 instruments and play methods to construct classifiers. It is very difficult to compare the experiments to the peer research, since different instruments were used in different study. Sound sources having significant pitch were chosen in the experiments. In each mix, sounds may come from different instrument families or same instrument family based on the extended Hornbostel and Sachs system [33]. Table 1 shows an example of a sound mix, where the second sound source is predominant, since the ratio to its original sound is 1. P/M represents the playing method.

**Table 1. An example of a sound mix - Group I.**

	2 <sup>nd</sup> Sound Source	1 <sup>st</sup> Sound Source
Pitch	B3	D#3
Timbre	Cello	Piano
Family	Chordophone	Chordophone
P/M	Bowed	Plucked
Ratio	1	1/4

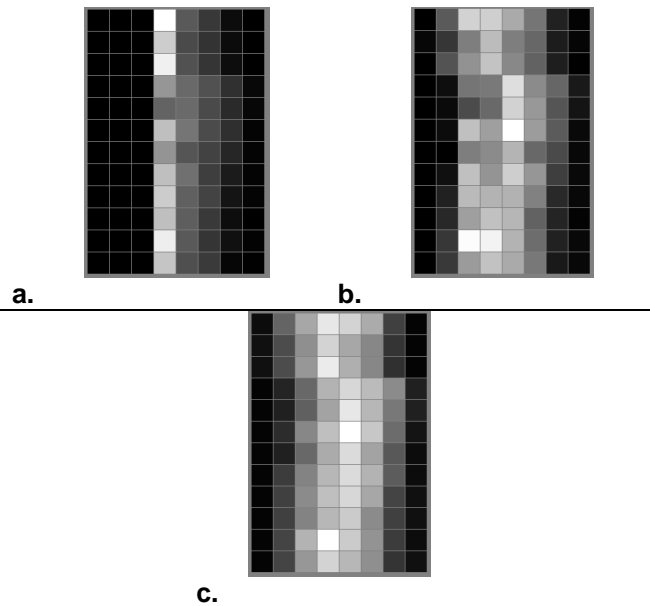
**Table 2. An example of a sound mix - Group II.**

	2 <sup>nd</sup> Sound Source	1 <sup>st</sup> Sound Source	3 <sup>rd</sup> Sound Source
Pitch	G4	A5	F2
Timbre	Viola	Bach Trumpet	Piano
Family	Chordophone	Aerophone	Chordophone
P/M	Bowed		Plucked
Ratio	1	1/4	1/8

Table 1 shows that an artificial sound mix contains two different sound sources, where sound volume in one of the sound sources were reduced to one fourth of the original

values to simulate pre-dominance. Table 2 shows that an artificial sound mix in Group II, containing three sound sources, where the sound volume of one sound source was reduced to one fourth and that of another sound source was reduced to one eighth.

With the clustering enhanced by pitch estimation device, 85 classifiers were automatically built; without the peak clustering technique, only 57 classifiers were constructed. Figure 5-a and 5-b show the classifier construction with and without peak clustering in the order from left to right, where each cell represents a classifier of a particular music note of a particular octave. The more frame objects used to construct the classifier, the brighter the cell. In Figure 5, the horizontal dimension is for octave in ascending order from left to right. The vertical dimension is for notes: A, A#, B, C, C#, D, D#, E, F, F#, G, and G# in the order from top to the bottom. Figure 5-c shows 96 manually categorized datasets by musicians, where each cell represents a dataset for a classifier of a particular note in a particular octave. The more sound objects in a dataset, the brighter its cell is. Please note that manually categorized dataset cannot be applied to construct classifiers, since the method needs to be applied for a classifier to learn as well as to estimate. The training data distribution from the clustering enhanced algorithm presented a more similar pattern to the categorization distribution based on human (musicians) perception than the none-clustering enhanced one. We also observed a significant improvement of pitch estimation precision especially for low pitches in the 2<sup>nd</sup> octave as well as high pitches in the 6<sup>th</sup> and 7<sup>th</sup> octaves.



**Figure 5. Dataset categorization for classifier construction: a. automatically by non-peak-clustering device; b. automatically by peak-clustering device; c. manually by musicians.**

The rest of section will discuss several tables about the precision of predominant pitch, timbre, and timber family estimation for different sound mix groups.

**Table 3. Predominant pitch estimation - Group I**

	Note (%)	Octave(%)
Aerophone	100.0	88.9
Chordophone	90.0	100.0
Average	95.0	94.5

**Table 4. Predominant pitch estimation - Group II**

	Note(%)	Octave(%)
Aerophone	100.0	91.7
Chordophone	100.0	83.3
Average	100.0	87.8

**Table 5. Predominant timbre estimation - Group I**

	Family(%)	Instrument(%)
Aerophone	100.0	66.7
Chordophone	100.0	100.0
Average	100.0	81.8

**Table 6. Predominant timbre estimation - Group II**

	Family(%)	Instrument(%)
Aerophone	100.0	100.0
Chordophone	100.0	55.6
Average	100.0	66.7

As shown in Table 3 and Table 4, the pitch estimation algorithm presented a considerably stable performance as the authors added more sources into the sound mix. Table 5 shows the precision of predominant sound source isolation for Group I, where each sound mix had two different sound sources. Table 6 shows the precision of timbre estimation for Group II, where each sound mix had three different sound sources. As more sound sources presented in the same channel of the recording, the timber estimation tended to be more difficult, due to the distraction of spectrum features among different sound sources.

## 4. CONCLUSIONS

The authors designed efficient pre-dominant harmonic sound sources separation algorithms to isolate sound sources by constructing dynamic classifiers of normalized harmonic partial relationships, which covered over 57 different music instruments with various play methods, while other music sound separation algorithms only have been proven upon very limited number of instruments. The system will cover more instruments, as the training database grows.

It is possible to estimate the most similar sound object in an audio database for a predominant sound source in a multi-timbre sound containing different pitches by a decision tree classifiers with harmonic sequential information, estimate the most similar frame by comparing a temporal feature in a feature database to perform spectrum subtraction, and repeats the steps until all the sound sources are subtracted from the multi-timbre sound. The decision tree classifier with temporal feature together can be used to improve the performance of instantaneous timbre estimation with accurate frame location estimation. Classification of harmonic peak sequence in different states separately provides efficient access to the feature database, and significantly improves accuracy.

Spectrum clustering aided maximum likelihood algorithm was robust for pitch estimation, spanning the big range of the frequencies, which western music instruments produce. Besides timber estimation, it may be also applied to research on melody, relationship between the human emotion factors and the music sound features, and music file annotation, etc.

## 5. ACKNOWLEDGMENTS

This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-0414815.

## REFERENCES

- [1] Balan, R. V., Rosca, J. P., Rickard, S. T. (2001). Robustness of Parametric Source Demixing in Echoic Environments, Proc. Int. Conf. on *Independent Component Analysis and Blind Source Separation (ICA)*, pp. 144-148.
- [2] Beauchamp, J. W., Maher, R.C., Brown, R. (1993). Detection of Musical Fundamental Frequency from Recorded Solo Performances. *94<sup>th</sup> Audio En. Society Convention*, preprint 3541, Berlin, March 16-19.
- [3] Berdahl, E., Burred, J.J. (2002) Moderne Methoden der Signal-analyse Abschlußbericht Grund-frequenz-analyse musikalischer Signale, Technische Universität Berlin - FG Kommunikation-swissenschaft, SoSe.
- [4] Brown, G.J., Cooke, M. P. (1994) Computational Auditory Scene Analysis, *Computer Speech and Language*, vol. 8, pp. 297-336.
- [5] Brown, J.C. (1992). Musical Fundamental Frequency Tracking Using a Pattern Recognition Method. *J. Acoust. Society Am.* 92(3), 1394-1402.
- [6] Brown, J.C. (1999). Musical Instrument Identification Using Pattern Recognition with Cepstral Coefficients as Features, *Journal of Acoustical Society of America*, 105(3), pp. 1933-1941.

- [7] Cardoso, J.F. (1998). Blind Source Separation: statistical principles, *Proceedings of the IEEE*, vol. 9, no. 10, pp.2009-2025.
- [8] Casey, M. A., Westner, A. (2000). Separation of Mixed Audio Sources by Independent Subspace Analysis, *Proc. International Computer Music Conference*, pp. 154-161.
- [9] Cherry, E. Collin. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *J. Acoustical Society of Am*, 24, pp. 975-979.
- [10] Cook, P.R., Morill, D., and Smith, J. O. (1998). An Automatic Pitch Detection and MIDI Control System For Brass Instruments. *J. Acoustical Society Am*, 92(4 pt. 2), 2429-2430.
- [11] Davies, M. E. (2002). Audio Source Separation, in *Mathematics in Signal Processing V*. Oxford Univ. Press.
- [12] Dziubinski, M., Dalka, P., Kostek, B (2005). Estimation of Musical Sound Separation Algorithm Effectiveness Employing Neural Networks, *Journal of Intelligent Information Systems*, 24(2/3), 133-158.
- [13] Eronen, A., Klapuri, A. (2000). Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. *Proceeding of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Plymouth, MA. pp. 753-756.
- [14] Fujinaga, I., McMillan, K. (2000). Real-Time Recognition of Orchestral Instruments, *Int. Comp. Music Conf.*, pp. 141-143.
- [15] Goto, M. (2000). A Robust Predominant-F0 Estimation Method for Real-Time Detection of Melody and Bass Line in CD Recordings. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, pp. II-757-760.
- [16] Hess, W. (1983). Fundamental frequency Determination of Speech Signals: Algorithms and Devices. Springer Berlin: Verlag, Tokyo: Heidelberg, New York.
- [17] Hyvarinen, A., Karhunen, J., Oja, E. (2001). Independent Component Analysis. John Wiley & Sons, 2001.
- [18] ISO/IEC JTC1/SC29/WG11 (2002). MPEG-7 Overview. (<http://mpeg.telecomitalia.com/standards/mpeg-7/mpeg-7.htm>)
- [19] Jensen, K. and Arnspang, J. (1999). Binary Decision Tree Classification of Musical Sounds, *the International Computer Music Conference*, Beijing, China.
- [20] Jourjine, A., Rickard, S., Yilmaz, O. (2000). Blind Separation of Disjoint Orthogonal Signals: Demixing N sources from 2 mixtures, *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, V-2985-2988.
- [21] Kaminskyj, I., Materka, A. (1995). Automatic Source Identification of Monophonic Musical Instrument Sounds, *IEEE International Conf. on Neural Networks*. 1, 189-194.
- [22] Klapuri, A. (1999). Wide-band Pitch Estimation for Natural Sound Sources with In-harmonicities. *106<sup>th</sup> Audio Eng. Society Convention*, Preprint 4906, Munich, May 8-11.
- [23] Martin, K.D., Kim, Y.E. 1998. Musical Instrument Identification: Pattern-Recognition Approach. *136th Meeting of Acoustical Society of America*, Norfolk, VA. 2pMU9.
- [24] Martin, K. D. (1999). Sound-Source Recognition: A Theory and Computational Model, Ph.D. Dissertation, MIT, Cambridge, MA.
- [25] McAulay, R., Quatieri, T. (1986). Speech Analysis/Synthesis Based on Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6), 744-754.
- [26] Ozerov, A., Philippe, P., Gribonval, R., Bimbot, F. "One microphone singing voice separation using source adapted models", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 90-93.
- [27] Quinlan, J.R. (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA.
- [28] Schroeder, M.R. (1968). Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement. *J. Acoust. Society Am.*, 43, 829-834.
- [29] Smith, J.O., Serra, X. (1987). PARSHL: An Analysis/Synthesis Program for Non Harmonic Sounds Based on a Sinusoidal Representation. In *Proc. Int. Comp. Music Conf.* (pp.290-297), Urbana-Champaign, Illinois.
- [30] Vincent, E., Gribonval, R. (2005). Construction d'estimateurs oracles pour la separation de sources, *Proc. 20th GRETSI Symp. on Signal and Image Processing*, 1245-1248.
- [31] Virtanen, T., Klapuri, A. (2000). Separation of Harmonic Sound Sources Using Sinusoidal Modeling. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Istanbul, Turkey.
- [32] Walmsley, P., Godsill, S.J., Rayner, P.J. (1999). Polyphonic Pitch Tracking Using Joint Bayesian Estimation of Multiple Frame Parameters. *IEEE Workshop on Applications of signal Processing to Audio and Acoustics*, 17<sup>th</sup>-20<sup>th</sup> October: New Paltz (NY).
- [33] Wiczorkowska, A., Ras, Z.W., Zhang, X., Lewis, R. (2007). Multi-way Hierarchic Classification of Musical Instrument Sounds, in *MUE 2007 Proceedings*, IEEE Computer Society, in Seoul, South Korea, 897-902.
- [34] Zhang, X., Marasek, K., Ras, Z. W. (2007). Maximum Likelihood Study for Sound Pattern Separation and Recognition, in *MUE 2007 Proceedings*, IEEE Computer Society, in Seoul, South Korea, 807-812.
- [35] Zhang, X. Ras, Z.W. (2006). Differentiated Harmonic Feature Analysis on Music Information Retrieval for Instrument Recognition, in *Proceeding of IEEE International Conference on Granular Computing*, May 10-12, Atlanta, Georgia, 578-581.
- [36] Zhang, X., Ras, Z.W., Dardzinska, A. (2008). Discriminant Feature Analysis for Music Timbre Recognition and Automatic Indexing, in *Mining Complex Data*, LNAI, Vol. 4944, Springer, 2008, 104-115.