Blind Music Timbre Source Isolation by Multi-resolution Comparison of Spectrum Signatures

Xin Zhang[1] , Wenxin Jiang[2], Zbigniew W. Ras[2,4,5], Rory Lewis[3]

[1] Univ. of North Carolina, Dept. of Math. and Comp. Science, Pembroke, NC 28372, USA
[2] Univ. of North Carolina, Dept. of Comp. Science, Charlotte, NC 28223, USA
[3] Univ. of Colorado, Dept. of Comp. Science, Colorado Springs, CO 80933, USA
[4] Polish-Japanese Institute of Information Technology, 02-008 Warsaw, Poland
[5] Polish Academy of Sciences, Institute of Comp. Science, 01-237 Warsaw, Poland
xin.zhang@uncp.edu, {wjiang3, ras}@uncc.edu, rlewis@eas.uccs.edu

Abstract. Automatic indexing of music instruments for multi-timbre sounds is challenging, especially when partials from different sources are overlapping with each other. Temporal features, which have been successfully applied in monophonic sound timbre identification, failed to isolate music instrument in multi-timbre objects, since the detection of the start and end position of each music segment unit is very difficult. Spectral features of MPEG7 and other popular features provide economic computation but contain limited information about timbre. Being compared to the spectral features, spectrum signature features has less information loss; therefore may identify sound sources in multi-timbre music objects with higher accuracy. However, the high dimensionality of spectrum signature feature set requires intensive computing and causes estimation efficiency problem. To overcome these problems, the authors developed a new multi-resolution system with an iterative spectrum band matching device to provide fast and accurate recognition.

Keywords: Blind Music Sound Sources Isolation, STFT (Short-Time Fourier Transform), Automatic Indexing, KNN, Spectral Features

1 Introduction

The rapid advances in computer storage and network techniques brought the emergency of huge multimedia repositories, where fast access to individual segment piece becomes more and more important in demands while manual indexing is a non-trivial work. Automatic indexing of music instruments in the same channel is one of the important subtasks.

A piece of digital music recording in a raw format contains some header information about the file and a huge sequence of sampling data of integers to represent the air fluctuations of sounds over time, where a typical sampling data rate is 44,100 per second for compact discs.

Features, such as MPEG-7 descriptors and other popular features, which are successfully applied in identifying music timbre in monophonic sounds, fail to isolate music source in multi-timbre or polyphonic objects, where multiple music instruments

are played at the same time. More so, temporal features are difficult to be applied in multi-timbre or polyphonic objects, since the detection of the start and end position of each music segment unit is very difficult while the partials are overlapping with each other (so-called a Cocktail Party Problem [6]).

Numerous methods for blind signal separation have been explored for a wide range of business domain spanning from finance to general biomedical signal processing. Filtering Techniques ([2], [3], [20]), ICA ([4], [7], [9]) and DUET [12] require different sound sources to be stored separately in multiple channels; therefore they are not suitable in isolating blind music sources in the same channel of the recordings. Most often, Factorial Hidden Markov Models (HMM [16]) work 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 [5] 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) [8]; therefore they are designed to deal with only harmonic sounds. For the purpose of interpolating the breaks in the sinusoidal component trajectories, numerous mathematical models have been explored: linear models [21], non-linear models such as high degree interpolation polynomials with cubic spine approximation model [8], 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. Kitahara et al. developed weights for features to minimize the influence of sound overlaps [13], which also assumes perfect fundamental frequency detection. Spectral features have been explored in peer research with traditional classifiers and proved a possible way to identify sound sources in multi-timbre music objects [11]. However, such features intuitively do not include sufficient information about sound wave behaviors along time. Also, when spectrum signatures are fed into classical classifiers, the order of frequency bins won't be taken into consideration. Therefore, the estimation accuracies of the traditional classifiers with only spectral features are normally not desirable. To overcome the problem, the authors developed a spectrum band matching device based on multi-resolution iterations to provide fast and accurate estimation based on an enlarged estimation range from the classifiers with relaxed confidence level for music instrument families.

2 Blind Music Timbre Source Isolation System

The authors developed a robust blind music sound source separation system with connection to a database of features extracted from a wide range of western music orchestral instruments, which consists of five major modules: a STFT converter with hamming window, a feature extraction engine, a K-Nearest-Neighbor classifier, an

iterative sound band matching device, and an FFT subtraction mechanism for the estimated predominant sound source.

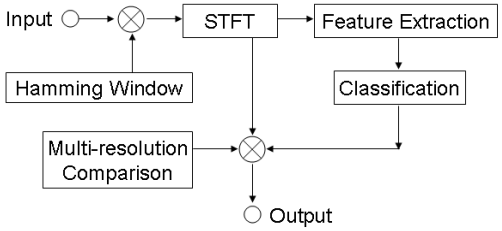

Fig. 1. System overview

The STFT converter divides a digital audio object into a sequence of frames, applies STFT transform to the mixed sample data of digital music data from time domain to frequency domain with a hamming window sliding evenly over time, and outputs NFFT (next-larger power-of-two number of samples of digital sound data from sampling window) discrete points.

The feature extraction engine calculates spectral features based on the spectrum information of the adjacent frames and stores them into a large music database used for training classifiers. In the next section, more details about those features will be presented.

The K-Nearest-Neighbor classifier takes in the flat spectrum features, constructs models, and estimates timbre categorization in terms of a series of machine understandable schemes.

An iterative sound band matching device is applied to further trim the bottom level of the tree models so that only the closely matched exemplary spectrum will remain, where each iteration rules out a certain amount of unlikely objects.

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 (Inverse discrete Fourier Transform) for each frame, and outputs resultant remaining signals into a new audio data file.

3 Feature Extraction

The authors developed a large database with spectral features and temporal attributes including popular features in this research area, such as MPEG7 spectral descriptors and Mel frequency cepstral coefficients, as well as some new temporal features.

Spectrum Centroid and Spread [1] The Audio Power Spectrum Centroid describes the center-of-gravity of a log-frequency power spectrum. Spectrum spread is defined as the Root Mean Square value of the deviation of the Log frequency power spectrum with respect to the gravity center in a frame. These two parameters economically indicate the pre-dominant frequency range.

Spectrum Flatness [1] describes the flatness property of the power spectrum within a frequency bin, which is ranged by edges. It is an array of aggregations in a

set of frequency bands, where frequency band is defined by two adjacent cutting edges with a quarter octave resolution spanning eight octaves.

Spectrum Basis Functions [1] are used to reduce the dimensionality of a group of adjacent frames of the normalized spectrum envelope in a log-arithmetic scale with a quarter-octave resolution by projecting from the space of 32 dimensions of frequency bands into a space of 10 dimensions with compact salient statistical information based on singular value decomposition.

Spectrum Projection Function [1] is computed by an inner product of the resultant low dimensional spectrum vector from the spectrum basis functions and the normalized spectrum envelope in a log-arithmetic scale. It is used to represent low-dimensional features of a spectrum after projection against a reduced rank basis of 10.

Predominant Harmonic Peaks [22] is an array of power spectrum coefficients of the local harmonic peaks in a normalized log-arithmetic scale based on the predominant fundamental frequency, where the first 28 of items are considered significant and therefore chosen as features in this research.

Harmonic Spectral Centroid [1] is computed as the average over the sound segment duration in the quasi-steady state of the instantaneous harmonic spectral centroid within a frame. The instantaneous harmonic spectral centroid is computed as the amplitude in a linear scale weighted mean of the harmonic peak of the spectrum.

Harmonic Spectral Spread [1] is computed as the average over the sound segment duration in the quasi-steady state of the instantaneous harmonic spectral spread within a frame. The instantaneous harmonic spectral spread is computed as the amplitude weighted standard deviation of the harmonic peaks of the spectrum with respect to the instantaneous harmonic spectral centroid.

Harmonic Spectral Variation [1] is defined as the mean value over the sound segment duration of the instantaneous harmonic spectral variation, which is calculated as the normalized correlation between the amplitude of the harmonic peaks of the current frame and the immediate previous frame.

Harmonic Spectral Deviation [1] is computed as the average over the sound segment duration of the instantaneous Harmonic Spectral Deviation in each frame, which is computed as the spectral deviation of the log amplitude components from a global spectral envelope.

Temporal Centroid [1] is calculated as the time average over the signal envelope.

Zero crossing [17], [19] counts the number of times that the signal sample data changes signs in a frame.

Roll-off is a measure of spectral shape, which is used to distinguish between voiced and unvoiced speech [14]. The roll-off is defined as the frequency below which a proportion (empirical value: 85%) of the accumulated magnitudes of the spectrum is concentrated.

Flux is used to describe spectral rate of change [17]. It is computed by the total difference between the magnitude of FFT points in a frame and its successive frame.

Mel frequency cepstral coefficients describe the spectrum according to the human perception system in the mel scale [15]. 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). The authors used the MFCC functions from the Julius software toolkit [1].

4 Classification

Numerous types of classifiers have been explored in timbre estimation by peer researchers, while so far there is no classifier, which is supreme in identifying all types of timbres in polyphonic or multi-timbre sounds among peer types of classifiers [23]. In this research, to explore the recognition rate of popular peer spectral features, decision tree was applied; while for spectrum signature features, K-Nearest-Neighbor algorithm was chosen for its fair performance with high dimensional feature sets (over 9,600 dimensions), where each frequency bin was treated as a feature. In case that accumulated error in a high dimensional space may bias the final estimation of timbre, we relaxed the confidence level, so that a group of possible candidates were collected as the output of the KNN classifiers. Further, a multi-resolution comparison device was applied to rule out unlikely candidates.

5 Multi-resolution Comparison

Searching for the closest matched pattern through high resolution of over eight thousands of FFT points by Euclidean distance may endanger the result by accumulated error as well as by the loss of order information along the frequency dimension. Actually, it is also opposite to the human visionary perception system. For example, when one recognizes a picture of the Eiffel Tower, does he or she checks from beam to beam assuming that beam is the atomic unit in the picture? No, on the contrary, most people would rather start from the outline shape, which is an abstract of details. In this research, authors started searching through vectors of aggregation of the frequency bins by an exponent order of resolution from low to high, where each round of comparison rules out a certain percentage of unlikely spectrum patterns as shown in the Figure 2.

In each round, the spectrum signature $V(\alpha)$ is computed by the following formula:

$$V_i^k(\alpha) = \left(\sum_{n=1}^{N_k(\alpha)} 10\log_{10}\frac{\chi_{i*N_k(\alpha)+n}}{\chi_{\max}-\chi_{\min}}\right) / N_k(\alpha) \quad (1)$$

where $\alpha \in \{3,4,5\}$ is the base, $V_i^k(\alpha)$ *is the* i^{th} feature in the k^{th} resolution level, χ is a vector of the power spectrum coefficients, and $N_k(\alpha)$ is computed by

$$N_k(\alpha) = \frac{M}{\alpha^k} \quad (2)$$

where M is the total number of FFT points. To limit the total number of iterations, Table 1 is used to show what values of k are allowed for each α; in each round/level k, 1 out of α^k points is chosen; α is used to yield even distribution of each resolution.

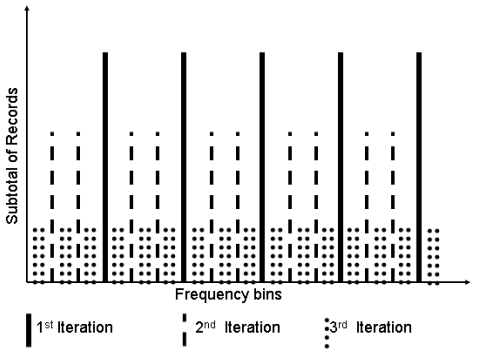

Fig. 2. Comparisons are applied based on iterative aggregation in resolution from low to high.

Table 1. The relationship between α and k. Positive sign means the adoption of the combination of α and k values in our experiments.

	k=1	k=2	k=3	k=4	k=5
α=3	+	+	+	+	+
α=4	+	+	+	+	
α=5	+	+	+		

Given a dataset D with S records in total, where $S >> M$, the time complexity C_α' of the spectrum signature matching strategy based on spectrum signature feature set can be represented by comparison cost, assuming that φ is the percentage of the total FFT that remains in the k^{th} resolution level.

$$C_\alpha' = M \cdot S \cdot \varphi^{8-\alpha} \sum_{k=0}^{8-\alpha} \left(\frac{1}{\alpha \cdot \varphi} \right)^k, \quad \alpha = 3, 4, 5 \tag{3}$$

In a flat spectrum signature comparison system, the time complexity C is $O(MS)$, while the comparison of C with C_α' for each base level α is listed in the following table (φ=50% was chosen in this example for the sake of simplicity):

Table 2. Percentagewise speed comparison between a new exponential multi-resolution spectrum signature matching device and a flat matching device.

	C_α'	C
α=3	8.55	100
α=4	12.11	100
α=5	20.30	100

Theoretically inferred by the above formula and data, the time complexity of spectrum signature matching strategy can be dramatically improved by the proposed multi-resolution spectrum matching system.

5 Experiments

In this research, the STFT experiments used a sampling window size of 0.12 second and a hop size of 0.04 second on music recording segments at the sampling rate of 44,100Hz, which is a typical value for compact disks. The training dataset contains 121790 spectrum signatures for the frames in the stable state of 3323 musical segment objects, which are played in the fourth octave C and originated from the MUMS (McGill University Masters Samples), assuming that similar results may be generated from music objects in other pitches. In real multimedia database, the data size of spectrum signatures will be in billions or trillions, as the musical segment objects were sampled every one second in short music sounds, of which the duration varies from around one to three seconds. Each spectrum signature contained 8192 FFT points. The training dataset included 26 music instruments: electric guitar, bassoon, oboe, b-flat clarinet, marimba, c trumpet, e-flat clarinet, tenor trombone, French horn, flute, viola, violin, English horn, vibraphone, accordion, electric bass, cello, tenor saxophone, b-flat trumpet, bass flute, double bass, alto flute, piano, Bach trumpet, tuba, and bass clarinet. The testing dataset consisted of 52 music recording pieces synthesized by Sound Forge sound editor [18], where each piece was played by two different music instruments.

The system was implemented in .NET C++ and MS SQLSERVER2005. The K Nearest Neighbor classifier package used in the experiments was from Microsoft SQLServer 2005. $K = 7$ was chosen empirically.

Two experiments were investigated to compare the efficiency and accuracy of the features for multi-timbre sounds: one was to check the accuracy of the popular peer features against the multi-resolution spectrum features; the other was to check the efficiency of multi-resolution spectrum signatures. In both experiments, accumulated confidence values were applied as votes for the top instrument candidates. In experiment I, we focused on the recognition rate instead of efficiency, since the peer spectral features contained much less dimensions of information than spectrum signatures; therefore the corresponding recognition results were fast and of low rate. In experiment II, linked lists were used to store the band coefficients for each tie of the resolution.

To compare the results with the traditional feature based classification strategy, five groups of spectral features (calculated for spectrum divided into 33 frequency bands) were extracted mainly from the MPEG-7 standard introduced in the previous section of Feature Extraction and fed into a set of decision tree classifiers for timbre estimation:

Group1: *Band Coefficients* = $\{b_n : 1 \le n \le 33\}$ – coefficients for Spectrum Flatness bands.

Group2: *Projections* = $\{p_n: 1 \le n \le 33\}$ – Spectrum Projection dimensions.

Group3: *MFCC* = $\{m_n: 1 \le n \le 13\}$ – Mel frequency cepstral coefficients.

Group4: *Harmonic Peaks* = $\{h_n: \ 1 \leq n \leq 28\}$ – harmonic partials of the predominant sound source.

Group5: Other Features include:

- Temporal Centroid,
- Log-arithmetic Spectral Centroid,
- Log-arithmetic Spectral Spread,
- Energy,
- Zero Crossings,
- Spectral Centroid,
- Spectral Spread,
- RollOff,
- Flux,
- Sum of the Spectrum Flatness band coefficients,
- Minimum, maximum, sum, distance, and standard deviation of the Spectrum Projection dimensions as well as of the Spectrum Basis dimensions, where distance represents a dissimilarity measure: distance of a matrix is calculated as the sum of absolute values of differences between each pair of elements on different rows and columns. Distance for a vector is calculated as the sum of dissimilarity (absolute difference of values) of every pair of coordinates in the vector.

The performance of our algorithm was measured using recognition rate R, calculated as the percentage of the correct estimations over the existing ones in the multi-timbral sound pieces.

Table 3. Music instrument recognition rate in experiment I

Experiment description	Recognition Rate (%)
Spectral features + decision tree	48.65
Flat spectrum features + KNN	82.43
α-base resolution spectrum features + KNN (α=3, 4, 5)	82.43

Table 3 shows that the multi-resolution spectrum features system with KNN classifiers had the same recognition rate as the flat one, which were both significantly better than the spectral features.

Table 4. Music instrument recognition efficiency in experiment II

Experiment description	Recognition Time (second)
Flat spectrum features + KNN	2560
α-base resolution spectrum features + KNN (α=3)	511
α-base resolution spectrum features + KNN (α=4)	524
α-base resolution spectrum features + KNN (α=5)	550

Table 4 shows that the multi-resolution spectrum features system significantly reduced the computing time to estimate the predominant music timbre in the music objects, which coincided the authors' theoretical derivation. The smaller the base, the more the iterations for the FFT points, therefore the faster the estimation. As the total

number of training objects in the multimedia database grows, the difference among recognition time of different resolution shall be further increased.

6 Conclusion

This research explored a new exponential multi-resolution spectrum signature matching device with KNN classifiers for blind music sound source isolation of multi-timbre musical objects. Temporal features were excluded in the experiments, since the detection of the start and end position of each multi-timbral music segment unit is very difficult and error prone. To compare the recognition rate, the authors developed two different training datasets: a spectral feature dataset and a spectrum signature feature dataset of multi-resolution. Traditional spectral features reduce data size for the limitation of input feature size of classic classifiers, but cause too much information loss for accurate music instrument detection. On the other hand, flat spectrum data is of high dimension and contains much more information, but does not suit most classic classifiers expect KNN. The authors designed a new algorithm with multi-resolution KNN and compared it with the peer spectral feature based algorithm. Overall, spectrum signature features were shown to provide significantly higher recognition rate for predominant music instrument then spectral features, as spectral features provided economic computation but contained not sufficient information for timbre recognition. Spectrum signature features with the multi-resolution matching device were proved same recognition rate as that with a flat matching device while the computing efficiency of the former system was much better than the latter one.

In the future, authors will explore the possibility to further improve the recognition rate of this exponential multi-resolution spectrum signature matching device with KNN classifiers by adding more carefully weighted new features, as the system can afford high dimensional dataset computing. On the other hand, feature selection algorithm may be applied to optimize the classification performance.

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

References

1. Akinobu, L. et al. Julius software toolkit. (http://julius.sourceforge.jp/en/)
2. Balan, R. V., Rosca, J. P., Rickard, S. T.: Robustness of parametric source demixing in echoic environments,in Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), pp. 144-148. (2001).
3. Brown, G. J., Cooke, M. P.: Computational auditory scene analysis, Computer Speech and Language, vol. 8, pp. 297-336. (1994).
4. Cardoso, J.F.: Blind source separation: statistical principles, Proceedings of the IEEE, vol. 9, no. 10, pp. 2009-2025. (1998).

5. Casey, M. A., Westner, A.: Separation of mixed audio sources by independent subspace analysis, in Proc. International Computer Music Conference. (ICMC), pp. 154-161. (2000).
6. Cherry, E.C.: Some Experiments on the Recognition of Speech, with One and with Two Ears. Journal of the Acoustical Society of America, 24, pp. 975-979. (1953).
7. Davies, M. E.: Audio source separation, in Mathematics in Signal Processing V. Oxford University Press. (2002).
8. Dziubinski, M., Dalka, P., Kostek, B.: Estimation of Musical Sound Separation Algorithm Effectiveness Employing Neural Networks, Journal of Intelligent Information Systems, 24(2/3), 133–158. (2005).
9. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons. (2001).
10. ISO/IEC JTC1/SC29/WG11. 2002. MPEG7 Overview. (http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm)
11. Jiang, W., Wieczorkowska, A., Ras, Z.W.: Music Instrument Estimation in Polyphonic Sound Based on Short-Term Spectrum Match", in "Data Mining: Theoretical Foundations and Applications", A.-E. Hassanien, A. Abraham, A. de Carvalho (Eds.), Studies in Computational Intelligence, Springer. (2009).
12. Jourjine, A. N., Rickard, S. T., Yilmaz, O.: Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures, in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. V-2985-2988. (2000).
13. Kitahara, T., Goto, M., Komatani, K., Ogata, T., Okuno, H. G.: Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps, EURASIP Journal on Advances in Signal Processing, Article ID 51979. (2007).
14. Lindsay, A. T., Herre, J.: MPEG7 and MPEG7 Audio—An Overview, J. Audio Engineering Society, Honolulu, Hawaii, July/Aug, vol.49, pp. 589–594. (2001).
15. Logan, B.: Mel Frequency Cepstral Coefficients for Music Modeling. In proceedings of 1st Annual International Symposium on Music Information Retrieval. (2000).
16. Ozerov,A., Philippe, P., Gribonval, R., Bimbot, F.: One microphone singing voice separation using source adapted models, in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 90-93. (2005).
17. Scheirer, E., Slaney, M.: Construction and Evaluation of a Robust Multi-feature Speech/Music Discriminator. In proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. (1997).
18. Sonic Foundry: Sound Forge. Software. (2003).
19. Tzanetakis, G., Cook, P.: Musical Genre Classification of Audio Signals, IEEE Transactions Speech and Audio Processing, July, vol. 10, pp. 293–302. (2002).
20. Vincent, E., Gribonval, R.: Construction d'estimateurs oracles pour la separation de sources, in Proc. 20th GRETSI Symposium on Signal and Image Processing, pp. 1245-1248. (2005).
21. Virtanen, T., Klapuri, A.: Separation of Harmonic Sound Sources Using Sinusoidal Modeling. In IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey. (2000).
22. Zhang, X., Marasek, K., Ras, Z.W.: Maximum Likelihood Study for Sound Pattern Separation and Recognition. In proceedings of International Conference on Multimedia and Ubiquitous Engineering, April 26-28, in Seoul, Korea. pp. 807-812. (2007)
23. Zhang, X., Ras, Z.W.: Analysis of Sound Features for Music Timbre Recognition. Invited paper, in proceedings of the IEEE CS International Conference on Multimedia and Ubiquitous Engineering, April 26-28, in Seoul, Korea, 3-8. (2007).