Discriminant Feature Analysis for Music Timbre Recognition and Automatic Indexing

Xin Zhang[1], Zbigniew W. Raś[1,3], and Agnieszka Dardzińska[2]

[1] Univ. of North Carolina, Dept. of Comp. Science, Charlotte, N.C. 28223, USA;
[2] Bialystok Technical Univ., Dept. of Comp. Science, ul. Wiejska 45a 15-351
Bialystok, Poland;
[3] Polish-Japanese Institute of Information Technology, ul. Koszykowa 86, 02-008
Warsaw, Poland;
e-mail: {xinzhang, ras, adardzin}@uncc.edu

Abstract. The high volume of digital music recordings in the internet repositories has brought a tremendous need for a cooperative recommendation system to help users to find their favorite music pieces. Music instrument identification is one of the important subtasks of a content-based automatic indexing, for which authors developed novel new temporal features and built a multi-hierarchical decision system S with all the low-level MPEG7 descriptors as well as other popular descriptors for describing music sound objects. The decision attributes in S are hierarchical and they include Hornbostel-Sachs classification and generalization by articulation. The information richness hidden in these descriptors has strong implication on the confidence of classifiers built from S. Rule-based classifiers give us approximate definitions of values of decision attributes and they are used as a tool by content-based Automatic Indexing Systems (*AIS*). Hierarchical decision attributes allow us to have the indexing done on different granularity levels of classes of music instruments. We can identify not only the instruments playing in a given music piece but also classes of instruments if the instrument level identification fails. The quality of *AIS* can be verified using precision and recall based on two interpretations: user and system-based [16]. *AIS* engine follows system-based interpretation.

1 Introduction

The state of art technologies in semantic web and computer storage boost the fast growing of music repositories throughout the internet, which in turn brought the need for intelligent search methods and efficient recommendation systems to help users to find their favorite music pieces.

Mining for knowledge in different representations of musical files (e.g., music recordings, MIDI files, and music notes) involves very different techniques. Research in MIDI files and music notes tackles problems in text mining. Digital recordings contain only sound signals unless manually labelled with semantic descriptions (e.g., author, title, and company). Knowledge mining in digital recordings requires prior retrieval of a large number of sound features from these musical sound signals. Timbre identification is one of the important subtasks for mining digital recordings, where timbre is a quality of sound that distinguishes one music instrument from another. Researchers in this area have investigated a number of acoustical features to build computational model for timbre identification. In this paper, authors focus on developing automated indexing solutions for digital recordings based on MIR (Music Information Retrieval) techniques of instruments and their types.

The real use of timbre-based grouping of music is very nicely discussed in [3]. Methods in research on automatic musical instrument sound classification go back to the last few years. We review these methods with respect to monophonic and polyphonic musical sounds.

For monophonic sounds, a number of acoustic features have been explored in [1], [4]. Some of them are quite successful for certain classes of sound data (monophonic, short, limited type of instruments). A digital multimedia file normally contains a huge amount of data, where subtle changes of sound amplitude in time can be critical for human perception system, thus the data-driven timbre identification process demands lots of information to be captured and also demands to describe the patterns among those subtle changes. Since after the dimensional approach to timbre description was proposed in [3], there is no standard parameterization used as a classification basis. Researchers in the area have explored a number of statistical parameters to describe patterns and properties of spectrums of music sounds to distinguish different timbre, such as Tristimulus parameters [14], [6], and irregularity [22], etc.

MPEG-7 standard provides a set of low-level temporal and spectral sound features where some of them are in a form of vector or matrix of a large size. Flattening and summarizing these features for traditional classifiers intuitively increases the number of features but losses some potentially useful information. Therefore, in this paper, authors have proposed a new set of features, sufficient in musical timbre signatures and suitable in format for machine learning classifiers. Authors compare them against popular features in the literature.

For polyphonic sounds, different methods have been investigated by various researchers, such as Independent Component Analysis (ICA) ([8], [21]), Factorial Hidden Markov Models (HMM) ([12], [19]), and Harmonic Sources Separation Algorithms ([2], [25], [9], [26]). ICA requires multiple channels of different sound sources. Most often, HMM works well for sound sources separation where fundamental frequency range is small and the variation is subtle. Harmonic Sources Separation Algorithms can be used to isolate sound sources within a single channel, where efficient solution in one channel can be intuitively applied to other channels and therefore facilitates more types of sound recordings (e.g., monochannel and stereo with two or more channels).

Our multi-hierarchical decision system is a database of about 1,000,000 musical instrument sounds, each one represented as a vector of approximately 1,100 features. Each instrument sound is labelled by a corresponding instrument. There are many ways to categorize music instruments, such as by playing methods, by instrument type, or by other generalization concepts [23]. Any categorization process is usually represented as a hierarchical schema which can be used by an automatic indexing system and a related cooperative Query Answering System (QAS) [7], [15], [17]. By definition, a cooperative QAS is relaxing a failing query with a goal to find its smallest generalization which does not fail. Two different hierarchical schemas [17] have been used as models of a decision attribute: Hornbostel-Sachs classification of musical instruments and classification of musical instruments by articulation. Each hierarchical classification represents a unique decision attribute, in a database of music instrument sounds, leading to a construction of a new classifier and the same to a different system for automatic indexing of music by instruments and their types [17], [28].

2 Audio Features in our Research

In their previous work, authors implemented aggregation [28] to the MPEG7 spectral descriptors as well as other popular sound features. This section introduces new temporal features and other popular features used to describe sound objects which we implemented in MIRAI database of music instruments [http://www.mir.uncc.edu]. The spectrum features have two different frequency domains: Hz frequency and Mel frequency. Frame size was carefully designed to be 120ms, so that the 0th octave G (the lowest pitch in our audio database) can be detected. The hop size is 40ms with a overlapping of 80ms. Since the sample frequency of all the music objects is 44,100Hz, the frame size is 5,292. A hamming window is applied to all STFT transforms to avoid jittering in the spectrum.

2.1 Temporal features based on pitch

Pitch trajectories of instruments behave very differently in time. The authors designed parameters to capture the power change in time.

Pitch Trajectory Centroid PC is used to describe the center of gravity of the power of the fundamental frequency during the quasi-steady state.

(1.) $PC = \frac{\sum_{n=1}^{length(P)}[\frac{n \cdot P(n)}{length(P)}]}{\sum_{n=1}^{length(P)} P(n)}$

where P is the pitch trajectory in the quasi-steady state, n is the n^{th} frame.

Pitch Trajectory Spread PS is the RMS deviation of the pitch trajectory with respect to its gravity center.

(2.) $PS = \sqrt{\frac{\sum_{n=1}^{length(P)}[(\frac{n}{length(P)} - PC)^2 \cdot P(n)]}{\sum_{n=1}^{length(P)} P(n)}}$

Pitch Trajectory Max Angle PM is an angle of the normalized power maximum vs. its normalized frame position along the trajectory in the quasi-steady state.

(3.) $PM = \frac{[\frac{MAX(P(n))-P(0)}{\frac{1}{length(P)}\cdot\sum_{n=1}^{length(P)} P(n)}]}{[\frac{F(n)-F(0)}{length(P)}]}$

where $F(n)$ is the position of n^{th} frame in the steady state.

Harmonic Peak Relation HR is a vector describing the relationship among the harmonic partials.

(4.) $HR = \frac{1}{m}\sum_{j=1}^{m}\frac{H_j}{H_0}$

where m is the total number of frames in the steady state, H_j is the j^{th} harmonic peak in the i^{th} frame.

2.2 Aggregation features

MPEG7 descriptors can be categorized into two types: temporal and spectral. The authors applied aggregation among all the frames per music object for all the following instantaneous spectral features.

MPEG7 Spectrum Centroid [29] describes the center-of-gravity of a log-frequency power spectrum. It economically indicates the pre-dominant frequency range. Coefficients under 62.5Hz have been grouped together for fast computation.

MPEG7 Spectrum Spread is the root of mean square value of the deviation of the Log frequency power spectrum with respect to the gravity center in a frame [29]. Like spectrum centroid, it is an economic way to describe the shape of the power spectrum.

MPEG7 Harmonic Centroid is computed as the average over the sound segment duration of the instantaneous harmonic centroid within a frame [29].

The instantaneous harmonic spectral centroid is computed as the amplitude in linear scale weighted mean of the harmonic peak of the spectrum.

MPEG7 Harmonic Spread is computed as the average over the sound segment duration of the instantaneous harmonic spectral spread of frame [29].

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.

MPEG7 Harmonic Variation is defined as the mean value over the sound segment duration of the instantaneous harmonic spectral variation [29].

The instantaneous harmonic spectral variation is defined as the normalized correlation between the amplitude of the harmonic peaks of two adjacent frames.

MPEG7 Harmonic Deviation is computed as the average over the sound segment duration of the instantaneous harmonic spectral deviation in each frame.

The instantaneous harmonic spectral deviation is computed as the spectral deviation of the log amplitude components from a spectral envelope.

MPEG7 Harmonicity Rate is the proportion of harmonics in the power spectrum. It describes the degree of harmonicity of a frame. It is computed by the normalized correlation between the signal and a lagged representation of the signal.

MPEG7 Fundamental Frequency is the frequency that best explains the periodicity of a signal. The ANSI definition of psycho-acoustical terminology says that "pitch is an auditory attribute of a sound according to which sounds can be ordered on a scale from low to high".

MPEG7 Upper Limit of Harmonicity describes the frequency beyond which the spectrum cannot be considered harmonic. It is calculated based on the power spectrum of the original and a comb-filtered signal.

Tristimulus and similar parameters describe the ratio of the amplitude of a harmonic partial to the total harmonic partials [26]. They are first modified tristimulus parameter, power difference of the first and the second tristimulus parameter, grouped tristimulus of other harmonic partials, odd and even tristimulus parameters.

Brightness is calculated as the proportion of the weighted harmonic partials to the harmonic spectrum [10].

(4.) $B = \frac{\sum_{n=1}^{N}[n \cdot A_n]}{\sum_{n=1}^{N} A_n}$

Transient, steady and decay duration. In this research, the transient duration is considered as the time to reach the quasi-steady state of fundamental frequency. At this duration the sound contains more timbre information than pitch information that is highly relevant to the fundamental frequency. Thus differentiated harmonic descriptors values in time are calculated based on the subtle change of the fundamental frequency [27].

Zero crossing counts the number of times that the signal sample data changes signs in a frame [20]

(5.) $ZC_j = 0.5 \sum_{n=1}^{N} \mid sign(s_j[n]) - sign(s_j[n-1]) \mid$

(6.) $sign(x)$ = [if $x \geq 0$ then 1, else -1]

where s_j is the n^{th} sample in the j^{th} frame, N is the frame size.

Spectrum Centroid describes the gravity center of the spectrum [24]

(7.) $C_j = \frac{\sum_{k=1}^{\frac{N}{2}} f(k) \cdot |X_j(k)|}{\sum_{k=1}^{\frac{N}{2}} |X_j(k)|}$

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

Roll-off is a measure of spectral shape, which is used to distinguish between voiced and unvoiced speech [11]. The roll-off is defined as the frequency below which C percentage of the accumulated magnitudes of the spectrum is concentrated, where C is an empirical coefficient.

Flux is used to describe the spectral rate of change [18]. It is computed by the total difference between the magnitude of the FFT points in a frame and its successive frame.

(8.) $F_j = \sum_{k=1}^{\frac{N}{2}}(\mid X_j(k) \mid - \mid X_{j-1}(k) \mid)^2$

2.3 Statistical parameters

In order to flatten the matrix data to suitable format for the classifiers, statistical parameters (e.g., maximum, minimum, average, distance of similarity, standard deviation) are applied to the power of each spectral band.

MPEG7 Spectrum Flatness describes the flatness property of the power spectrum within a frequency bin, which is ranged by edges in the corresponding formula (see [29]). The value of each bin is treated as an attribute value in the database. Since the octave resolution in our research is 1/4, the total number of bands is 32.

MPEG7 Spectrum Basis Functions are used to reduce the dimensionality by projecting the spectrum from high dimensional space to low dimensional space with compact salient statistical information (see [29]).

Mel Frequency Cepstral Coefficients describe the spectrum according to the human perception system in the Mel scale. They are computed by grouping the STFT points of each frame into a set of 40 coefficients by a set of 40 weighting curves with logarithmic transform and a discrete cosine transform (DCT).

2.4 MPEG7 temporal descriptors

The temporal descriptors in MPEG7 [29] have been applied directly into the feature database. **MPEG7 Spectral Centroid** is computed as the power weighted average of the frequency bins in the power spectrum of all frames in a sound segment with Welch method. **MPEG7 Log Attack Time** is defined as the logarithm of the time duration between the time when the signal starts to the time it reaches its stable part, where the signal envelope is estimated by computing the local mean square value of the signal amplitude in each frame. **MPEG7 Temporal Centroid** is calculated as the time average over the energy envelope.

3 Discriminant Analysis for Feature Selection

Logistic regression model is a popular statistical approach of analyzing multinomial response variables. 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 discrete responses and the explanatory attributes. It has been widely used to investigate the relationship between decision attribute and conditional attributes, using the most economical model. An ordinal response logit model has a form:

(9.) $(\frac{Pr(Y=i|x)}{Pr(Y=k+1|x)}) = \alpha_i + \beta_i \cdot x$, $i = 1, 2, ..., k$

where the $k + 1$ possible responses have no natural ordering and α_1,..., α_k are k intercept parameters, β_1,..., β_k are k vectors of parameters, and Y is the response. For details, see [5]. The system fits a common slopes cumulative model which is a parallel lines regression model based on the cumulative probabilities of the response categories. The significance of an attribute is calculated with the likelihood ratio or chi-square difference test by the Fisher's Score algorithm. A final model is selected, where adding another variable would not improve the model significantly.

4 Experiments

The authors used a subset of their feature database [http://www.mir.uncc.edu] containing 1,569 music recording sound objects of 74 instruments. The authors discriminated instrument types on different levels of a classification tree. The tree consists of three levels: the top level (e.g., aerophone, chordophone, and idiophone), the second level (e.g., lip-vibrated, side, reed, composite, simple, rubbed, shaken, and struck), and the third level (e.g., piano, violin, and flute). All classifiers were 10-fold cross validation with a split of 90% training and 10% testing. We used WEKA for all classifications and SAS LOGISTIC procedure for discriminant analysis. In each experiment, a 99% confidence level was used. Feature extraction was implemented in .NET C++ with connection to MS SQL Server. In LISP notation, we used the following *Music Instrument Classification Tree*:

(Instrument(Aerophone(Lip-vibrated (-,-,-), Side(-,-), Reed(-,-)),
Chordophone(Composite, Simple), Idiophone(Rubbed(-), Shaken(-,-),
Struck(-,-))))

For classification on the first level in the music instrument family tree, the selected feature set was stored in List I: {PeakRelation8, PeakRelation16, PeakRelation24, MPEGFundFreq, MPEGHarmonicRate, MPEGULHarmonicity, MPEGHarmoVariation, MPEGHarmoDeviation, MPEGFlat3, MPEGFlat8, MPEGFlat18, MPEGFlat30, MPEGFlat36, MPEGFlat46, MPEGFlat55, MPEGFlat56, MPEGFlat66, MPEGFlat67, MPEGFlat76, MPEGFlat77, MPEGFlat83, MPEGFlat85, MPEGFlat94, MPEGFlat96, MPEGSpectrumCentroid, MPEGTC, MPEGBasis59, MPEGBasis200, TristimulusRest, ZeroCrossing, MFCCMaxBand1, MFCCMaxBand3, MFCCMaxBand5, MFCCMaxBand6, MFCCMaxBand7, MFCCMaxBand8, MFCCMaxBand10, MFCCMaxBand13, MFCCMinBand1, MFCCMinBand13, PitchSpread, MaxAngle}. Experiment was also performed on the rest of features after List I was removed from the whole feature set, which was stored in List II. In the table below, "All" stands for all the attributes used for classifier construction.

Table 1 shows the precisions of the classifiers constructed with selected features at the family level. After the less significant features, elected by the logistic model, have been removed, the group of List I slightly improved the precision for aerophone instruments. However, the selected significant feature group (List I) significantly outperformed in precision for aerophone instruments and in recall for both chordophone and aerophone instruments.

	Precision			*Recall*		
Class	List I	All	List II	List I	All	List II
Idiophone	87.00%	91.10%	95.10%	82.10%	91.40%	94.80%
Chordophone	86.80%	91.30%	88.50%	88.60%	88.90%	84.70%
Aerophone	91.50%	91.30%	87.30%	91.80%	93.50%	90.90%

Table 1. Results of three groups of features at the top level of the music family tree

For classification at the second level in the music instrument family tree, the selected feature set was stored in List I: {PeakRelation8, PeakRelation16, PeakRelation30, MPEGFundFreq, MPEGHarmonicRate, MPEGULHarmonicity, MPEGHarmoDeviation, MPEGFlat3, MPEGFlat11, MPEGFlat14, MPEGFlat18, MPEGFlat22, MPEGFlat26, MPEGFlat36, MPEGFlat44, MPEGFlat46, MPEGFlat58, MPEGFlat67, MPEGFlat81, MPEGFlat82, MPEGFlat83, MPEGFlat85, MPEGFlat93, MPEGFlat94, MPEGFlat95, MPEGSpectrumCentroid, MPEGTC, MPEGBasis50, MPEGBasis57, MPEGBasis59, MPEGBasis69, MPEGBasis73, MPEGBasis116, MPEGBasis167, MPEGBasis206, Tristimulus1, TristimulusRest, TristimulusBright, ZeroCrossing, SpectrumCentroid2, RollOff, MFCCMaxBand1, MFCCMaxBand3, MFCCMaxBand4, MFCCMaxBand6, MFCCMaxBand7, MFCCMaxBand9, MFCCMinBand2, MFCCMinBand5, MFCCMinBand10, MFCCMinBand13, MFCCAvgBand10, MFCCAvgBand11, PitchSpread, MaxAngle}. Experiment was also performed on List II obtained by removing List I from the whole feature set.

	Precision			*Recall*		
Class	List I	All	List II	List I	All	List II
$Lip-Vibrated$	83.80%	84.40%	77.30%	84.70%	88.80%	82.30%
Side	74.30%	73.20%	66.40%	75.70%	64.00%	64.00%
Reed	77.10%	78.30%	70.50%	78.40%	80.10%	70.50%
Composite	84.50%	86.20%	84.90%	86.70%	84.30%	83.90%
Simple	71.20%	74.10%	72.20%	67.20%	80.00%	72.80%
Rubbed	85.30%	82.10%	75.00%	78.40%	86.50%	73.00%
Shaken	79.20%	91.00%	89.50%	64.80%	92.00%	87.50%
Struck	78.20%	86.30%	85.40%	80.40%	79.00%	77.60%

Table 2. Results of three groups of features at the second level of the music family tree

Table 2 shows the precisions of the classifiers constructed with the selected features, all features, and the rest of the features after selection at the second level of the instrument family tree. After the less significant features, elected by the logistic model, have been removed, the group of List I improved the precision for side and rubbed instruments and recall for the side, composite, and struck instruments. Also, the selected significant feature group (List I) significantly outperformed in precision for lip-vibrated, side, reed, and rubbed instruments and in recall for all the types except for simple and shaken instruments.

For classification at the third level in the music instrument family tree, the selected feature set was stored in List I: { MPEGTristimulusOdd, MPEGFundFreq, MPEGULHarmonicity, MPEGHarmoVariation, MPEGFlatness6, MPEGFlatness14, MPEGFlatness27, MPEGFlatness35, MPEGFlatness43, MPEGFlatness52, MPEGFlatness63, MPEGFlatness65, MPEGFlatness66, MPEGFlatness75, MPEGFlatness76, MPEGFlatness79, MPEGFlatness90, MPEGFlatness91, MPEGSpectrumCentroid, MPEGSpectrumSpread, MPEGBasis41, MPEGBasis42, MPEGBasis69, MPEGBasis87, MPEGBasis138, MPEGBasis157, MPEGBasis160, MPEGBasis170, MPEGBasis195, TristimulusBright, TristimulusEven, TristimulusMaxFd, ZeroCrossing, SpectrumCentroid2, Flux, MFCCMaxBand2, MFCCMaxBand3, MFCCMaxBand6, MFCCMaxBand7, MFCCMaxBand9, MFCCMaxBand10, MFCCMinBand1, MFCCMinBand2, MFCCMinBand3, MFCCMinBand6, MFCCMinBand7, MFCCMinBand10, MFCCAvgBand1, MFCCAvgBand12, SteadyEnd, Length}. Experiment was also performed on List II obtained by removing List I from the whole feature set.

	Precision			*Recall*		
Class	List I	All	List II	List I	All	List II
Flute	92.90%	67.70%	70.40%	89.70%	72.40%	65.50%
TubularBells	86.70%	60.00%	52.40%	72.20%	66.70%	61.10%
Tuba	85.70%	81.80%	85.70%	90.00%	90.00%	90.00%
ElectricBass	83.10%	87.50%	89.10%	80.60%	83.60%	85.10%
Trombone	80.60%	80.60%	76.30%	69.20%	82.10%	74.40%
Marimba	79.20%	89.50%	90.00%	71.40%	89.50%	86.50%
Piano	78.50%	82.40%	83.00%	81.60%	78.40%	74.40%
FrenchHorn	78.00%	83.70%	82.90%	87.70%	88.90%	84.00%
BassFlute	77.40%	75.50%	71.20%	68.30%	61.70%	61.70%
AltoFlute	76.70%	82.80%	78.60%	79.30%	82.80%	75.90%
DoubleBass	75.40%	60.80%	60.00%	75.40%	54.40%	52.60%
Piccolo	74.50%	69.20%	62.00%	71.70%	67.90%	58.50%
CTrumpet	72.00%	68.90%	69.10%	83.10%	78.50%	72.30%
Violin	71.00%	75.00%	77.10%	78.00%	72.70%	76.50%
Oboe	70.30%	71.00%	35.90%	81.30%	68.80%	43.80%
Vibraphone	69.30%	91.40%	85.70%	73.20%	90.10%	93.00%
Bassoon	68.80%	66.70%	45.50%	61.10%	55.60%	27.80%
Cello	67.00%	63.20%	63.50%	61.50%	62.50%	68.80%
Saxophone	66.70%	51.70%	53.60%	46.70%	50.00%	50.00%

Table 3. Results of three groups of features at the third level of the music family tree

Table 3 shows statistics of the precisions of the classifiers constructed with the selected features, all features, and the rest of the features after selection in the bottom level of the family tree for some instruments in the experiment. The overall accuracy of all the features was slightly better than that of the selected features. The computing time for List I, All, and List II is 7.33, 61.59, and 54.31 seconds respectively.

5 Conclusion and future work

A large number of attributes is generated in a table during fattening the features into a single value attributes for classical classifiers by statistical and other feature design methods. Some of the derived attributes may not significantly contribute to the classification models, or sometimes may distract the classification. In the light of the results from the experiments, we conclude that attributes have different degree of influence on the classification performance for different instrument families. The new temporal features related to harmonic peaks significantly improved the classification performance when added into the database with all other features. However, the new features were not suitable to replace the MPEG7 harmonic peak related features and Tristimulus parameters as the logistic studies shows. We also noticed that classifications at a higher level of granularity tended to use more features for correct prediction than those at the lower level. This may especially benefit a cooperative query answering system to choose suitable features for classifiers at different levels.

6 Acknowledgements

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

References

1. Balzano, G.J. (1986) What are musical pitch and timbre? Music Perception, an interdisciplinary Journal, Vol. 3, 297-314
2. Bay, M. and Beauchamp, J.W. (2006) Harmonic source separation using prestored spectra, ICA 2006, LNCS 3889, 561-568
3. Bregman, A.S. (1990) Auditory scene analysis, the perceptual organization of sound, MIT Press
4. Cadoz, C. (1985) Timbre et causalite, unpublished paper, Seminar on Timbre, Institute de Recherche et Coordination Acoustique/Musique, Paris, France, April 13-17
5. Cessie, S., Houwelingen, J.C. (1992) Ridge Estimators in Logistic Regression, Applied Statistics, Vol. 41, No. 1, 191-201
6. Fujinaga, I., McMillan, K. (2000) Real time recognition of orchestral instruments, Proceedings of the International Computer Music Conference, 141-143
7. Gaasterland, T. (1997) Cooperative answering through controlled query relaxation, in IEEE Expert, Vol. 12, No. 5, 48-59

8. Kinoshita, T., Sakai, S., and Tanaka, H. (1999) Musical sound source identification based on frequency component adaptation, in Proceedings of IJCAI Workshop on Computational Auditory Scene Analysis (IJCAI-CASA '99), Stockholm, Sweden, July-August, 18-24
9. Kitahara, T., Goto, M., Komatani, K., Ogata, T., Okuno, H.G. (2007) Instrument identification in polyphonic music: feature weighting to minimize influence of sound overlaps, in EURASIP Journal on Advances in Signal Processing, No. 1, 155-155
10. Lewis, R., Zhang, X., Ras, Z.W. (2007) Knowledge discovery based identification of musical pitches and instruments in polyphonic sounds, Journal of Engineering Applications of Artificial Intelligence, Elsevier, Vol. 20, No. 5, 637-645
11. Lindsay, A. T., and Herre, J. (2001) MPEG-7 and MPEG-7 Audio-An Overview, J. Audio Eng. Soc., vol.49, July/Aug, 589-594
12. Ozerov,A., Philippe, P., Gribonval, R., and Bimbot, F. (2005) One microphone singing voice separation using source adapted models, in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 90-93
13. Pawlak, Z., (1991) Information systems - theoretical foundations, in Information Systems Journal, Vol. 6, 205-218
14. Pollard, H.F., Jansson, E.V. (1982) A tristimulus Method for the specification of Musical Timbre, Acustica, Vol. 51, 162-171
15. Ras, Z.W., Dardzińska, A. (2006) Solving Failing Queries through Cooperation and Collaboration, Special Issue on Web Resources Access, (Editor: M.-S. Hacid), in World Wide Web Journal, Springer, Vol. 9, No. 2, 173-186
16. Ras, Z.W., Dardzińska, A., Zhang, X. (2007) Cooperative Answering of Queries based on Hierarchical Decision Attributes, CAMES Journal, Polish Academy of Sciences, Institute of Fundamental Technological Research, Vol. 14, No. 4, 729-736
17. Ras, Z.W., Zhang, X., Lewis, R. (2007) MIRAI: Multi-hierarchical, FS-tree based Music Information Retrieval System, (Invited Paper), Proceedings of RSEISP 2007, M. Kryszkiewicz et al. (Eds), LNAI, Vol. 4585, Springer, 80-89
18. Scheirer, E. and Slaney, M. (1997) Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)
19. Smith, J.O. and Serra, X. (1987) PARSHL: An Analysis/Synthesis Program for Non Harmonic Sounds Based on a Sinusoidal Representation, Proc. Int. Computer Music Conf., Urbana-Champaign, Illinois, 290-297
20. Tzanetakis, G., Cook, P. (2002) Musical Genre Classification of Audio Signals, IEEE Trans. Speech and Audio Processing, July, Vol. 10, 293-302
21. Vincent, E. (2006) Musical source separation using time-frequency source priors, IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 1, 91-98
22. Wieczorkowska, A. (1999) Classification of musical instrument sounds using decision trees, Proceedings of the 8th International Symposium on Sound Engineering and Mastering, ISSE 1999, 225-230
23. Wieczorkowska, A., Ras, Z., Zhang, X., Lewis, R. (2007) Multi-way Hierarchic Classification of Musical Instrument Sounds, in Proceedings of the International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), Seoul, South Korea, IEEE Computer Society, 897-902
24. Wold, E., Blum, T., Keislar, D., and Wheaton, J.(1996) Content-Based Classification, Search and Retrieval of Audio, IEEE Multimedia, Fall, 27-36
25. Zhang, X., Marasek, K., Ras, Z. W. (2007) Maximum likelihood study for sound pattern separation and recognition, in Proceedings of the IEEE CS International

Conference on Multimedia and Ubiquitous Engineering (MUE 2007), April 26-28, Seoul, Korea, 807-812

26. Zhang, X., Ras, Z.W. (2007) Sound isolation by harmonic peak partition for music instrument recognition, in the Special Issue on Knowledge Discovery, Fundamenta Informaticae Journal, IOS Press, Vol. 78, No. 4, 613-628
27. Zhang, X., Ras, Z.W. (2006) Differentiated Harmonic Feature Analysis on Music Information Retrieval For Instrument Recognition, Proceeding of IEEE International Conference on Granular Computing, May 10-12, Atlanta, Georgia, 578-581
28. Zhang, X., Ras, Z.W. (2007) Analysis of sound features for music timbre recognition, in Proceedings of the IEEE CS International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), April 26-28, Seoul, Korea, 3-8
29. ISO/IEC JTC1/SC29/WG11 (2002) MPEG-7 Overview,
http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm