A Study on Feature Analysis for Musical Instrument Classification

Da Deng[†], Christian Simmermacher, Stephen Cranefield
Dept. of Information Science, University of Otago
P O Box 56, Dunedin, New Zealand
[†] ddeng@infoscience.otago.ac.nz

Abstract

In tackling data mining and pattern recognition tasks, finding a compact but effective set of features has often been found to be a crucial step in the overall problem-solving process. In this paper we present an empirical study on feature analysis for classical instrument recognition, using machine learning techniques to select and evaluate features extracted from a number of different feature schemes. It is revealed that there is significant redundancy between and within feature schemes commonly used in practice. Our results suggest that further feature analysis research is necessary in order to optimize feature selection and achieve better results for the instrument recognition problem.

1 Introduction

Music data analysis and retrieval has become a very popular research field in recent years. The advance of signal processing and data mining techniques has led to intensive study on content-based music retrieval [1][2], music genre classification [3][4], duet analysis [2], and most frequently, on musical instrument detection and classification (e.g., [5][6][7][8]).

Instrument detection techniques can have many potential applications. For instance, detecting and analyzing solo passages can lead to more knowledge about different musical styles and be further utilized to provide a basis for lectures in musicology. Various applications for audio editing, audio and video retrieval or transcription can be supported. An overview of audio information retrieval has been presented by Foote [9] and extended by various authors [2][10]. Other applications include playlist generation [11], acoustic environment classification [12, 13], and using audio feature extraction to support video scene analysis and annotation [14].

One of the most crucial aspects of instrument classification, is to find the right feature extraction scheme. During the last few decades, research on audio signal processing has focused on speech recognition, but few features can be directly applied to solve the instrument classification problem. New methods are being investigated so as to achieve semantic interpretation of low-level features extracted by audio signal processing methods. For example, a framework of low-level and high-level features given in the MPEG-7 multimedia description standard [15] can be used to create application-specific

description schemes. These can be used to annotate music with a minimum of human supervision for the purpose of music retrieval.

In this paper, we present a study on feature extraction and selection for instrument classification using machine learning techniques. Features were first selected by ranking and other schemes, data sets of reduced features were generated, and their performance in instrument classification was further tested with a few classifiers using cross validations. A number of feature schemes were considered based on human perception, cepstral features, and the MPEG-7 audio descriptors. The performance of the feature schemes was assessed first individually, and then in combination with each other. We also used dimension reduction techniques so as to gain insight on the right dimensionality for feature selection. Our aim was to find differences and synergies between different feature schemes and test them with various classifiers, so that a robust classification system could be built. *Features extracted from different feature schemes were ranked and selected, and a number of classification algorithms were employed and managed to achieve good accuracies in three groups of experiments: instrument family classification, individual instrument classification, and classification of solo passages.*

Following this introduction, Section 2 reviews the recent relevant work on musical instrument recognition and audio feature analysis. Section 3 outlines the approach we adopted in tackling the problem of instrument classification, including feature extraction schemes, feature selection methods, and classification algorithms used. Experiment settings and results based on the proposed approach are then presented in Section 4. We summarize the findings and conclude the paper in Section 5.

2 Related Work

Various feature schemes have been proposed and adopted in the literature of instrument sound analysis. On top of the adopted feature schemes, different computational models or classification algorithms have been employed for the purposes of instrument detection and classification.

Mel-frequency cepstral coefficients (MFCC) features are commonly employed not only in speech processing, but also in music genre classification and instrument classification. Marques and Moreno [5] built a classifier that can distinguish between eight instruments with 70% accuracy using Support Vector Machines (SVM). Eronen [6] assessed the performance of MFCC features and spectral and temporal features such as amplitude envelope and spectral centroids for instrument classification. The Karhunen-Loeve transform was conducted to decorrelate the features, and *k*-nearest neighbors (*k*-NN) classifiers were used with their performance assessed through cross validation. The results favoured MFCC features, and violin and guitar were among the most poorly recognized instruments.

The MPEG-7 audio framework targets standardization of the extraction and description of audio features [15][16]. The sound description of MPEG-7 audio features was assessed by Peeters et al. [17] based on their perceived timbral similarity. It was concluded that combinations of the MPEG-7 descriptors can be reliably applied in assessing the similarity of musical sounds. Xiong et al. [12] compared MFCC and MPEG-7 audio features for the purpose of sports audio classification, adopting hidden Markov models (HMM) and a number of classifiers such as *k*-NN, Gaussian mixture models, AdaBoost, and SVM. Kim et al. [10] examined the use of HMM-based classifiers trained on MPEG-7 based audio descriptors to solve audio classification problems such as speaker recognition and sound classification.

Brown et al. [18] conducted a study on identifying four instruments of the woodwind family. Features used were cepstral coefficients, constant Q transform, spectral centroid, autocorrelation coefficients. For classification, a scheme using Bayes decision rules was adopted. The recognition rates based on the

feature sets varied from 79%-84%.

Agostini et al. [7] extracted spectral features for timbre classification, and the performance was assessed over SVM, *k*-NN, canonical discriminant analysis, and quadratic discriminant analysis, with the first and last being the best. Compared with the average 55.7% correct tone classification rate achieved by some conservatory students, it was argued that computer-based timbre recognition can exceed human performance at least for isolated tones.

Essid et al. [8] processed and analyzed solo musical phrases from ten instruments. Each instrument was represented by fifteen minutes of audio material from various CD recordings. Spectral features, audio spectrum flatness, MFCC, and derivatives of MFCC were used as features. SVM yielded an average result of 76% for 35 features.

Livshin and Rodet [19] evaluated the use of monophonic phrases for detection of instruments in continuous recordings of solo and duet performances. The study made use of a database with 108 different solos from seven instruments. A set of 62 features (temporal, energy, spectral, harmonic, and perceptual) was proposed and subsequently reduced by feature selection. The best 20 features were used for realtime performance. A leave-one-out cross validation using a *k*-NN classifier gave an accuracy of 85% for 20 features and 88% for 62 features. Benetos et al. [20] adopted branch-and-bound search to extract a 6-feature subset from a set of MFCC, MPEG-7, and other audio spectral features. A non-negative matrix factorization algorithm was used to develop the classifiers, gaining an accuracy of 95.2% for six instruments.

Kostek [2] studied the classification of twelve instruments played under different articulations. She used multilayer neural networks trained on wavelet and MPEG-7 features. It was found that a combination of these two feature schemes can significantly improve the classification accuracy to a range of 55% - 98%, with an average of about 70%. Misclassifications occurred mainly within each instrument family (woodwinds, brass, and strings). A more recent study by Kaminskyj et al. [21] dealt with isolated monophonic instrument sound recognition using k-NN classifiers. Features used include MFCC, constant Q transform spectrum frequency, Root mean square (RMS) amplitude envelop, spectral centroid, and multidimension scaling analysis trajectories. These features underwent principal component analysis (PCA) to be reduced to a total dimensionality of 710. k-NN classifiers were then trained under different hierarchical schemes. A leave-one-out strategy was used, yielding an accuracy of 93% in instrument recognition, and 97% in instrument family recognition.

Some progress has been made in musical instrument identification for polyphonic recordings. Eggink and Brown [22] presented a study on the recognition of five instruments (flute, oboe, violin and cello) in accompanied sonatas and concertos. Gaussian-mixture-model classifiers were employed on features reduced by PCA. The classification performance on a variety of data resources ranged from 75% to 94%, while misclassification occurred mostly for flute and oboe (both classified as violin).

With the emergence of many audio feature schemes, feature analysis and selection has been gaining more research attention recently. A good introduction on feature selection was given in Guyon and Elisseef [23], outlining the methods of correlation modelling, selection criteria, and the general approaches of using filters and wrappers. Yu and Liu [24] discussed some generic methods such as information gain (IG) and symmetric uncertainty (SU), where an approximation method for correlation and redundancy analysis was proposed based on using SU as the correlation measure. Grimaldi et al. [25] evaluated selection strategies such as IG and gain ratio (GR) for music genre classification. Livshin and Rodet [19] used linear discriminant analysis to repeatedly find and remove the least significant feature, until a subset of 20 features was obtained from the original 62 feature types. The reduced feature set gave an

average classification rate of 85.2%, very close to that of the complete set.

Benchmarking is still an open issue that remains unresolved. There are very limited resources available for benchmarking, so direct comparison of these various approaches is not possible. Most studies have used recordings digitized from personal or institutional CD collections. McGill University Master Samples (http://www.music.mcgill.ca/resources/mums/html/mums.html) have been used in some studies [7][22][21], while the music samples from the MIS Database from UIOWA (http//:theremin.music.uiowa.edu/) were also widely used [18][6][22][20].

3 Feature Analysis and Validation

3.1 Instrument categories

Traditionally, musical instruments are classified into four main categories or families: string, brass, woodwind, and percussion. For example, violin is a typical string instrument, oboe and clarinet belong to the woodwind category, horn and trumpet are brass instruments. Piano is usually classified as a percussion instrument. Sounds produced by these instruments bear different acoustic attributes. A few characteristics can be obtained from these sound envelopes, including attack (the time from silence to amplitude peak), sustain (the time length in maintaining level amplitude), decay (the time the sound fades from sustain to silence), and release (the time of the decay from the moment the instrument stops playing). To achieve accurate classification of instruments, more complicated features need to be extracted.

3.2 Feature Extraction for instrument classification

Because of the complexity of modelling instrument timbre, various feature schemes have been proposed through acoustic study and pattern recognition research. One of our main intentions is to investigate the performance of different feature schemes as well as find a good feature combination for a robust instrument classifier. Here, we use three different extraction methods, namely, perception-based features, MPEG-7 based features, and MFCC. The first two feature sets consist of temporal and spectral features, while the last is based on spectral analysis. These features, 44 in total, are listed in Table 1. Among them the first sixteen are perception-based features, the next seven are MPEG-7 descriptors, and the last 26 are MFCC features.

3.2.1 Perception-based features

To extract perception-based features, music sound samples are segmented into 40ms frames with 10ms overlap. Each frame signal was analysed by 40 band-pass filters centered at Bark scale frequencies. The following are some important perceptual features that are used in this study:

- *zero-crossing rate* (ZCR), an indicator for the noisiness of the signal, often used in speech processing applications:

$$\text{ZCR} = \frac{\sum_{n=1}^{N} |sign(F_n) - sign(F_{n-1})|}{2N} \tag{1}$$

  where N is the number of samples in the frame, and F_n the value of the n-th sample of a frame.

Table 1. Feature Abbreviations and Descriptions

#	Abbr.	Description	Scheme
1	ZC	Zero Crossings	Perception-based
2-3	ZCRM, ZCRD	Mean and standard deviation of ZC Ratios	
4-5	RMSM, RMSD	Mean and standard deviation of RMS	
6-7	CentroidM, CentroidD	Mean and standard deviation of Centroid	
8-9	BandwidthM, BandwidthD	Mean and standard deviation of Bandwidth	
10-11	FluxM, FluxD	Mean and standard deviation of Flux	
12	HC	Harmonic Centroid Descriptor	MPEG-7
13	HD	Harmonic Deviation Descriptor	
14	HS	Harmonic Spread Descriptor	
15	HV	Harmonic Variation Descriptor	
16	SC	Spectral Centroid Descriptor	
17	TC	Temporal Centroid Descriptor	
18	LAT	Log-Attack-Time Descriptor	
19-44	MFCC*k*M, MFCC*k*D	Mean and standard deviation of the first 13 linear MFCCs	MFCC

- the *Root-mean-square* (RMS), which summarizes the energy distribution in each frame and channel over time:

$$\text{RMS} = \sqrt{\frac{\sum_{n=1}^{N} F_n^2}{N}} \tag{2}$$

- *Spectral centroid*, which measures the average frequency weighted by sum of spectrum amplitude within one frame:

$$\text{Centroid} = \frac{\sum_{k=1}^{K} P(f_k) f_k}{\sum_{k=1}^{K} P(f_k)} \tag{3}$$

  where f_k is the frequency in the k-th channel, the number of channels is K=40, and $P(f_k)$ the spectrum amplitude on the k-th channel.

- *Bandwidth*, also referred to as centroid width, shows the frequency range of a signal weighted by its spectrum:

$$\text{Bandwidth} = \frac{\sum_{k=1}^{K} |Centroid - f_k)| P(f_k)}{\sum_{k=1}^{K} P(f_k)} \tag{4}$$

- *Flux*, which represents the amount of local spectral change, calculated as the squared difference between the normalized magnitudes of consecutive spectral distributions:

$$\text{Flux} = \sum_{k=2}^{K} |P(f_k) - P(f_{k-1})|^2 \tag{5}$$

These features were extracted from multiple segments of a sample signal and it is the mean value and the standard deviation that are used as the feature values for each music sample.

3.2.2 MPEG-7 timbral features

Instruments have unique properties which can be described by their harmonic spectrums and their temporal and spectral envelopes. The MPEG-7 audio framework [15] endeavours to provide a complete feature set for the description of harmonic instrument sounds. We consider in this work only two classes of timbral descriptors in the MPEG-7 framework: Timbral Spectral and Timbral Temporal. These include seven feature descriptors: Harmonic Centroid (HC), Harmonic Deviation (HD), Harmonic Spread (HS), Harmonic Variation (HV), Spectral Centroid (SC), Log-Attack-Time (LAT), and Temporal Centroid (TC). The first five belong to the Timbral Spectral feature scheme, while the last two belong to the Timbral Temporal scheme. Note that the SC feature value is obtained from the spectral analysis of the entire sample signal, so it is similar but different from the CentroidM of the perception-based features. CentoidM is aggregated from the centroid feature analysed from short segments within a sample.

3.2.3 MFCC features

To obtain MFCC features, a signal needs to be transformed from frequency (Hertz) scale to mel scale:

$$mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{6}$$

The mel scale has 40 filter channels. The first extracted filterbank output is a measure of power of the signal, and the following 12 linear spaced outputs represent the spectral envelope. The other 27 log-spaced channels account for the harmonics of the signal. Finally a discrete cosine transform converts the filter outputs to give the MFCCs. Here, the mean and standard deviation of the first thirteen linear values are extracted for classification.

3.3 Feature Selection

Feature selection techniques are often necessary to optimize the feature set used for classification. This way, redundant features are removed from the classification process and the dimensionality of the feature set is reduced, so as to save computational cost and defy the "curse of dimensionality" that impedes the construction of good classifiers [23]. To assess the quality of a feature used for classification, a correlation-based approach is often adopted. In general, a feature is good if it is relevant to the class concept but is not redundant given the inclusion of other relevant features. The core issue is modelling the correlation between two variables or features. Based on information theory, a number of indicators can be developed to rank the features by their correlation to the class. Relevant features will yield a higher correlation.

Given a pre-discretized feature set, the 'noisiness' of the feature X can be measured as the entropy, defined as:

$$H(X) = -\sum_i P(x_i)\log_2 P(x_i), \tag{7}$$

where $P(x_i)$ is the prior probability for the *i*-th discretized value of X. The entropy of X after observing another variable Y is then defined as

$$H(X|Y) = -\sum_j P(y_j) \sum_i (P(x_i|y_j) \log_2 P(x_i|y_j)), \tag{8}$$

The Information Gain (IG) [26], indicating the amount of additional information about X provided by Y, is given as

$$\text{IG}(X|Y) = H(X) - H(X|Y) \tag{9}$$

IG itself is symmetrical, i.e., IG$(X|Y)$ =IG$(Y|X)$, but in practice it favours features with more values [24].

The gain ration method (GR) normalizes IG by an entropy term:

$$\text{GR}(X|Y) = \frac{\text{IG}(X|Y)}{H(Y)} \tag{10}$$

A better measure is defined as the symmetrical uncertainty [27]:

$$\text{SU} = 2\frac{\text{IG}(X|Y)}{H(X) + H(Y)} \tag{11}$$

SU compensates for IG's bias toward features with more values and restricts the value range within $[0, 1]$.

Despite a number of efforts previously made using the above criteria [25][24], there is no golden rule for the selection of features. In practice, it is found that the performance of the selected feature subsets is also related to the choice of classifiers for pattern recognition tasks. The wrapper-based approach [28] was therefore proposed, using a classifier combined with some guided search mechanism to choose an optimal selection from a given feature set.

3.4 Feature analysis by dimension reduction

To get a reference level for deciding how many features are sufficient for problem solving, one can use standard dimension reduction or multidimension scaling (MDS) techniques such as PCA and Isomap [29] to assess an embedding dimension of the high-dimensional feature space. PCA projects high-dimensional data into low-dimension space while preserving the maximum variance. Naturally it is optimal for data compression but has also been found rather effective in pattern recognition tasks such as face recognition and handwriting recognition. The Isomap algorithm calculates the geodesic distances between points in a high-dimensional observation space, and then conducts eigenanalysis of the distance matrix. As the output, new coordinates of the data points in a low-dimensional embedding are obtained that best preserve their intrinsic geodesic distances. In this study, we used PCA and Isomap to explore the sparseness of the feature space and examine the residuals of the chosen dimensionality so as to estimate at least how many features should be included in a subset. The performance of the selected subsets was then compared with that of the reduced and transformed feature space using MDS.

3.5 Feature validation via classification

Feature combination schemes generated from the selection rankings were then further assessed using classifiers and cross-validated. The following classification algorithms were used in this study: *k*-NN,

an instance-based classifier weighted by the reciprocal of distances [30]; Naive Bayes classifier employs Bayesian models in the feature space; and SVM, which is a statistical learning algorithm and has been widely used in many classification tasks.

4 Experiment

4.1 Experiment settings

We tackled the music instrument classification problem in two stages: 1) instrument type classification using samples of individual instruments, and 2) direct classification of individual instruments.

A number of utilities were used for feature extraction and classification experiments. The perception-based features were extracted using the IPEM Toolbox [31]. The Auditory Toolbox [32] was used to extract MFCC features. The MPEG-7 audio descriptor features were obtained using an implementation by Casey [33]. Various algorithms implemented in Weka (Waikato Environment for Knowledge Analysis) [34] were used for feature selection and classification experiments.

Samples used in the first experiment were taken from the previously mentioned UIOWA MIS collection. The collection consists of 761 single instrument files from 20 instruments which cover the dynamic range from pianissimo to fortissimo and are played bowed or plucked, with or without vibrato, depending on the instrument. All samples were recorded in the same acoustic environment (anechoic chamber) under the same conditions. We realize that this is a strong constraint and our result may not generalize to a complicated setting such as live recordings of an orchestra. To explore the potential of various feature schemes for instrument classification in live solo performance, solo passage music samples were collected from music CD recordings from private collections and the University of Otago library.

In general, the purposes of these experiments is to test the performance of the feature schemes, evaluate the features using feature selection, and also test the performance of different classifiers.

4.2 Instrument family classification

4.2.1 Feature ranking and selection

We first simplified the instrument classification problem by grouping the instruments into four families: piano, brass, string and woodwind. For this four-class problem, the best 20 features of the three selection methods are shown in Table 2. All of them indicate that Log-Attack-Time (LAT) and Harmonic Deviation (HD) are the most relevant features. The following features have nearly equal relevance. It is important to note that the standard deviations of the MFCCs are predominantly present in all three selections. Also the measures of the centroid and bandwidth, as well as the deviation of flux, zero crossings and mean of RMS can be found in each of them. These selections are different from the best 20 features selected by Livshin and Rodet [19], where MPEG-7 descriptors were not considered. However, they also included bandwidth (spectral spread), MFCC, and Spectral Centroid.

Classifiers were then used to assess the quality of feature selection. A number of algorithms, including Naive Bayes, *k*-NN, multilayer perceptrons (MLP), radial basis function (RBF), and SVM, were compared on classification performance based on 10-fold cross validation. Among these, the Naive Bayes classifiers employed kernel estimation during training. A plain k-NN classifier was used here with $k = 1$. SVM classifiers were built using sequential mimimal optimisation, with RBF kernels and a complexity value of 100, with all attributes standardized. Pairwise binary SVM classifiers were trained

Table 2. Feature ranking for single tones.

Rank	IG		GR		SU		SVM
	Feature	*Value*	*Feature*	*Value*	*Feature*	*Value*	*Feature*
1	LAT	0.8154	LAT	0.5310	LAT	0.4613	HD
2	HD	0.6153	HD	0.5270	HD	0.3884	FluxD
3	FluxD	0.4190	MFCC2M	0.3230	BandwidthM	0.2267	LAT
4	BandwidthM	0.3945	MFCC12D	0.2970	FluxD	0.2190	MFCC3D
5	MFCC1D	0.3903	MFCC4D	0.2700	RMSM	0.2153	MFCC4M
6	MFCC3D	0.381	BandwidthM	0.2660	MFCC1D	0.2084	ZCRD
7	RMSM	0.3637	RMSM	0.2640	MFCC4M	0.1924	MFCC1M
8	BandwidthD	0.3503	MFCC13D	0.2580	MFCC11D	0.1893	HC
9	MFCC4M	0.3420	MFCC2D	0.2450	MFCC3D	0.1864	MFCC9D
10	MFCC11D	0.3125	MFCC11D	0.2400	BandwidthD	0.1799	ZC
11	ZCRD	0.3109	MFCC7D	0.2350	MFCC2M	0.1784	RMSM
12	CentroidD	0.2744	FluxD	0.2290	MFCC4D	0.1756	CentroidD
13	MFCC8D	0.2734	MFCC1D	0.2240	MFCC7D	0.1710	MFCC9M
14	MFCC6D	0.2702	MFCC4M	0.2200	MFCC12D	0.1699	BandwidthM
15	MFCC7D	0.2688	CentroidM	0.2150	ZCRD	0.1697	MFCC5D
16	ZC	0.2675	SC	0.2110	CentroidD	0.1653	SC
17	MFCC4D	0.2604	MFCC5M	0.2090	CentroidM	0.1610	MFCC12D
18	CentroidM	0.2578	CentroidD	0.2080	MFCC13D	0.1567	MFCC7M
19	MFCC10M	0.2568	HC	0.1950	SC	0.1563	MFCC2M
20	MFCC10D	0.2519	MFCC1M	0.1910	MFCC8D	0.1532	MFCC6M

Table 3. Classifier performance of the instrument families.

Feature Scheme	*k*-NN	Naive Bayes	SVM	MLP	RBF
All 44	95.75%	86.5%	97.0%	95.25%	95.0%
Best 20	94.25%	86.25%	95.5%	93.25%	95.5%
Best 10	90.25%	86.25%	94.25%	91.0%	87.0%
Best 5	89.5%	81.0%	91.75%	86.75%	84.5%

Table 4. Performance of classifiers trained on the "Selected 17" feature set.

Classifier	1NN	Naive Bayes	SVM	MLP	RBF
Performance	96.5%	88.25%	92.75%	94%	94%

for this multi-class problem, with between 10 and 80 support vectors created for each SVM. The structure of MLP is automatically defined in the Weka implementation, and each MLP is trained over 500 epochs with a learning rate of 0.3 and a momentum of 0.2.

To investigate the redundancy of the feature set, we used the Information Gain filter to generate reduced feature sets of the best 20, best 10, and best 5 features respectively. Other choices instead of IG were found to produce similar performance and hence were not considered further. The performance of these reduced sets was compared with the original full set with all 44 features. The results are given in Table 3.

These can be contrasted with results presented in Table 4, where 17 features were selected using a rank search based on SVM attribute evaluation and the correlation-based CfsSubset scheme implemented in Weka. This feature set, denoted as "Selected 17", includes CentroidD, BandwidthM, FluxD, ZCRD, MFCC[2-6]M, MFCC10M, MFCC3/4/6/8D, HD, LAT, and TC. It is noted that TC contributes positively to the classification task, even though it is not among the top 20 ranked features. Here the classification algorithms take similar settings as those used to generate the results shown in Table 3. The performance of the "Selected 17" feature set is very close to that of the full feature set. Actually the k-NN classifier performs even slightly better with the reduced feature set.

4.2.2 Evaluation of feature extraction schemes

Since the k-NN classifier produced similar performance in much less computing time compared with SVM, we further used 1-NN classifiers to assess the contribution from each individual feature scheme and improvements achieved through scheme combinations. Apart from combining the schemes two by two, another option was also considered, picking the top 50% ranked attributes from each feature scheme, resulting in a 21-dimension composite set, called 'Top 50% mix'. The results are presented in Table 5. Besides overall performance, classification accuracy on each instrument type is also reported.

From these results, it can be seen that among the individual feature subsets, MFCC outperforms both IPEM and MPEG-7. This is different from the finding of Xiong et al. [12] that MPEG-7 features give better results than MFCC for the classification of sports audio scenes such as applause, cheering, and music etc. The difference is however marginal (94.73% vs 94.60%). Given that the scope of our study is much narrower, this should not be taken as a contradiction. Indeed, some researchers also found more favourable results using MFCC instead of MPEG-7 for instrument classification [10][8].

In terms of average performance of combination schemes listed in Table 5, the MFCC+MPEG-7

Table 5. Performance (%) in classifying the 4 classes (10 CV)

Feature Sets	Brass	Woodwind	String	Piano	Overall
MFCC (26)	99	90	89	95	93.25
MPEG-7 (7)	90	62	76	99	81.75
IPEM (11)	93	63	81	100	84.25
MFCC+MPEG-7 (33)	98	92	91	100	95.25
MFCC+IPEM (37)	98	89	94	98	94.75
IPEM+MPEG-7(18)	93	76	85	100	88.5
Top 50% mix (21)	95	89	88	100	93
Best 20	97	88	92	100	94.25
Selected 17	97	94	95	100	96.5

set shows the best results, while the MPEG-7+IPEM set with 18 features has the poorest result. It is observed that the inclusion of MFCC is most beneficial for woodwind and string families, while the inclusion of the MPEG-7 seems to boost the performance on piano and woodwind. Generally, the more features that are included, the better the performance. However, between 33, 37 and 44 features the difference is almost negligible. It is interesting to note that the 'Selected 17' feature set produces very good performance. The top 50% mix set produces a performance as high as 93%, which is slightly worse than that of 'best 20' probably due to the fact that the selection is not done globally among all features. All these results, however, clearly indicate that there is strong redundancy within the feature schemes.

In terms of accuracy on each instrument type, the piano can be classified by most feature sets rather accurately. The MPEG-7 and IPEM sets have problems in identifying woodwind instruments, with which MFCC can cope very well. Combining MFCC with other feature sets can boost the performance on 'woodwind' significantly. The MPEG-7 set does not perform well on string instruments either, however, a combination with either MFCC or IPEM can effectively enhance the performance. These results suggest that these individual feature sets are quite complementary to each other. On the other hand, the good performance of the selected feature set clearly indicates that there is high redundancy among the three basic feature schemes.

4.2.3 Dimension reduction

Overall, when the total number of included features is reduced, the classification accuracy decreases monotonically. However, it is interesting to see from Table 3 that even with five features only, the classifiers achieved a classification rate around 90%. In order to interpret this finding, we used PCA and Isomap to reduce the dimensionality of the full feature set. The two methods report similar results. The normalized residuals of the extracted first 10 components using these methods are shown in Figure 1. The 3-D projection of the Isomap algorithm, generated by selecting the first three coordinates from the resulting embedding, is shown in Figure 2. For both methods the residual falls under 0.5% after the 4th component, although the dropping reported by Isomap is more significant. This suggests that the data manifold of the 44-dimensional feature space may have an embedded dimension of 4 or 5 only.

As a test, the first five principal components (PC) of the complete feature set were extracted, resulting in weighted combinations of MFCC, IPEM and MPEG-7 features. A 1-NN classifier trained with the five PCs reports an average accuracy of 88.0% in a 10-fold cross validation, very close to that of the

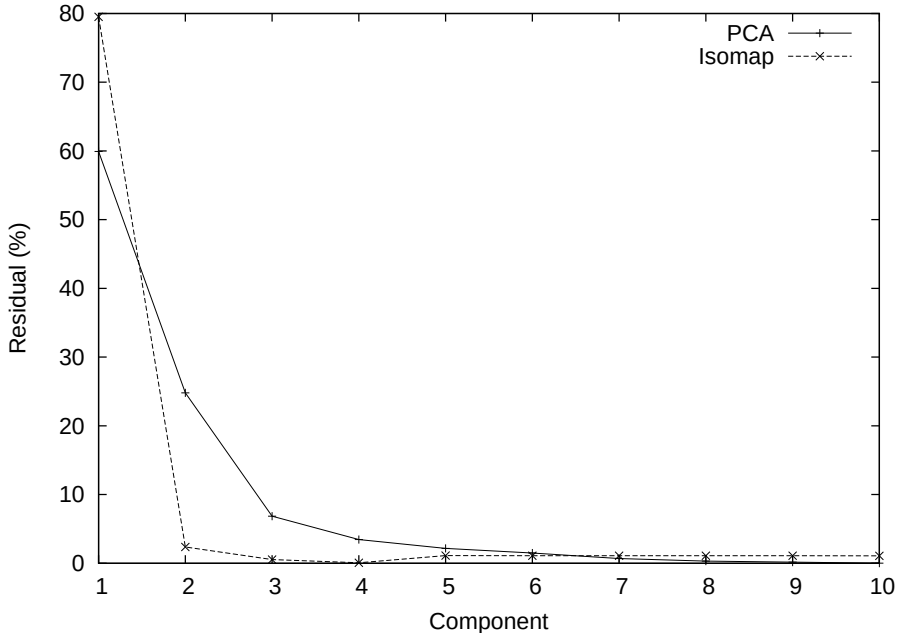

Figure 1. Dicree diagram of the reduced components. The x axis gives the component number, and the y axis gives the relevant normalized residual (in %). Only ten components are shown.

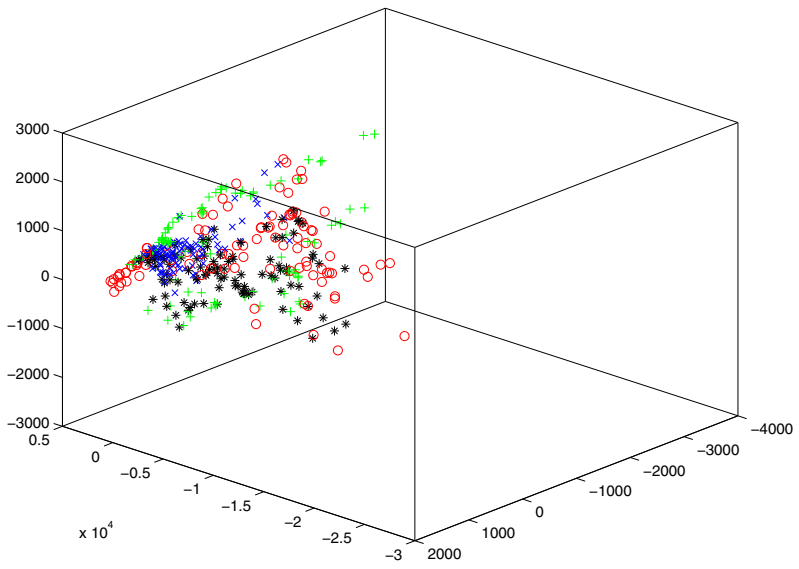

Figure 2. 3-D embedding of the feature space. There are 400 instrument samples, each with its category labelled: $\times$ - 'piano', $\circ$ - 'string', $+$ - 'brass', $*$ - 'woodwind'. The three axes correspond to the transformed first 3 dimensions generated by Isomap.

"Best 5" selection given in Table 3. This further confirms that there is strong redundancy within and between the three feature schemes.

4.3 Instrument Classification

4.3.1 Individual instrument sound recognition

Table 6. Confusion matrix for all 20 instruments with 10-fold CV.

Instrument	Classified As																			
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
a=piano	19	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
b=tuba	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c=trumpet	0	0	19	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d=horn	0	0	0	19	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
e=ttrombone	0	0	0	0	18	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
f=btrombone	0	0	0	0	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g=violin	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0
h=viola	0	0	0	0	1	2	1	18	0	0	0	0	0	0	0	1	0	1	0	1
i=bass	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	1	0	1	0	0
j=cello	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0
k=sax	0	0	0	0	0	0	0	0	0	0	8	1	0	0	0	0	0	0	0	1
l=altosax	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	2	0	0	0
m=oboe	0	1	0	0	0	1	0	1	0	1	0	0	6	0	0	0	0	0	0	0
n=bassoon	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
o=flute	0	0	0	0	0	0	0	0	1	0	0	0	0	0	7	1	0	0	1	0
p=altoflute	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7	2	0	0	0
q=bflute	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	8	0	0	0
r=bclarinet	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	6	0	0
s=bbclarinet	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	8	0
t=ebclarinet	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	2	0	0	0	5

Next, all 20 instruments were directly distinguished from each other. Here we chose to use 1-NN classifiers as they work very fast and give almost the same accuracies as compared to SVM. A feature selection process was conducted, using correlation-based subset selection on attributes searched by SVM evaluation. This resulted in a subset of 21 features, including LAT, FluxM, ZCRD, HD, CentroidD, TC, HC, RMSD, FluxD, and 12 MFCC values. The confusion matrix for individual instrument classification is given in Table 6. Instrument *a* is piano, and instruments *b-f* belong to the brass type, *g-j* the string type, and *k-t* the woodwind type.

The overall average classification accuracy is 86.9%. The performance in general is quite satisfactory, especially for piano and string instruments. Only one out 20 piano samples was wrongly classified (as oboe). Among string instruments, the most significant errors occurred for viola samples, with an accuracy of 18/25=72%. Classification errors in the woodwind category mainly occurred within itself, having only sporadic cases of wrong classification as instruments of other families. The woodwind instruments have the lowest classification accuracy compared with other instruments, but this may relate

to the limited number of woodwind data samples in the current data set. The worst classified instrument is E^b clarinet. There is also a notable confusion between alto flute and bass flute.

4.3.2 Classification of solo phrases

Finally, a preliminary experiment on solo phrases was conducted. For this experiment one representative instrument of each instrument type was chosen. These were: trumpet, flute, violin, and piano. To detect the right instrument in solo passages, a classifier was trained on short monophonic phrases. Ten second long solo excerpts from CD recordings were tested on this classifier. The problem here is that the test samples were recorded with accompaniment, thus are often polyphonic in nature. Selecting fewer and clearly distinguishable instruments for the trained classifier helps to make the problem more addressable. It is assumed that an instrument is playing dominantly in the solo passages. Thus, its spectral characteristics will probably be the most dominant and the features derived from the harmonic spectrum are assumed to work. Horizontal masking effects will probably the most crucial problem to tackle. Overlapping tones could mask the attack and decay time.

The samples for the four instruments were taken from live CD recordings. Passages of around ten seconds' length were segmented into two second phrases with 50% overlap. The amount of music samples was basically balanced across the four instrument types, as seen in Table 7. A change to shorter one second segments for training and testing showed similar results but with a tendency to lower recognition rates. The trumpet passages sometimes have multiple brass instruments playing. The flutes are accompanied by multiple flutes, a harp or a double bass, and the violin passages are sometimes flute and string accompanied.

Table 7. Data sources for the solo phrase experiment.

Trumpet	9 min / 270 samples
Piano	10.6 min / 320 samples
Violin	10 min / 300 samples
Flute	9 min / 270 samples
Total	38.6 min / 1160 samples

The same SVM-based feature selection scheme used before searched out 19 features for this task. These included: 8 MFCC values (mainly means), 5 MPEG-7 features (HD, HS, HV, SC), and 4 perception-based features (CentroidM, FluxM, ZCRD, and RMSM. An average accuracy of 98.4% was achieved over four instruments using 3-NN classifiers with distance weighting. The Kappa statistics is reported as 0.98 for 10-fold cross validation, which suggests that the classifier stability is very strong. The confusion matrix is shown in Table 8. Numbers shown are percentage. The largest classification errors occurred with flute being classified as piano.

Here again, MFCC was shown to be dominant in classification. To achieve a good performance, it is noted that the other two feature schemes also contributed favourably.

4.4 Discussion

The scopes of some current studies and performance achieved are listed in Table 9, where the number of instruments, classification accuracies (in %) of instrument family classification and individual instrument classification are listed. It can be seen that our results are better or comparable with those obtained

Table 8. Confusion matrix for intrument recognition in solo passages (performance in %).

Instrument	Classified As			
	piano	trumpet	violin	flute
piano	100	0	0	0
trumpet	0.4	99.6	0	0
violin	0.3	0.3	98.7	0.7
flute	3.7	0	1.5	94.8

by other researchers. However, it is noted that the number of instruments included is different, and the data sources are different despite the fact that most of these included the UIOWA sample set. The exact validation process used may be different as well. For instance, we used 10-fold cross validation, while Kaminskj and Czaszejko [21] and others used leave-one-out.

Paired with a good performance level, the feature dimensionality of our approach is relatively low with the selected feature sets having less than or around 20 dimensions. On the other hand, Eggink and Brown [22] used the same UIOWA sample collection but a different feature scheme with 90 dimensions, reporting an average recognition rate of only 59% on five instruments (flute, clarinet, oboe, violin and cello). Livshin and Rodet [19] used 62 features and selected the best 20 for real-time solo detection. Kaminskj and Czaszejko [21] used 710 dimensions after PCA. In our study, a 5-dimension set after PCA can also achieve a good classification accuracy. A notable work is by Benetos et al. [20], where only 6 features are selected. However, there are only six instruments included in their study and the scalability of the feature selection needs to be further assessed.

Table 9. Performance of instrument classification compared with that of ours.

Work	no. of instruments	Family classification	Individual classification
Eronen [6]	29	77	35
Martin and Kim [35]	14	90	70
Agostini et al. [7]	27	81	70
Kostek [2]	12	-	70
Kaminskyj and Czaszejko [21]	19	97	93
Benetos et al. [20]	6	-	95.2
This work	20	96.5	86.9

As for classification of solo passages, it is hard to make direct comparison as no common benchmarks have been accepted and researchers used various sources including performance CDs [8, 19]. With more benchmark data becoming available in the future, it is our intention to further assess the feature schemes and feature selection methods employed in this study.

5 Conclusion

In this paper, we presented a study on feature extraction and evaluation for the problem of instrument classification. The main contribution is that we investigated three major feature extraction schemes, analyzed them using feature selection measures based on information theory, and assessed their performance using classifiers undergoing cross validation.

For experiments on monotone music samples, a publicly available data set was used so as to allow for the purpose of benchmarking. Feature ranking measures were employed and these produced similar feature selection outputs, which basically aligns with the performance obtained through classifiers. The MPEG-7 audio descriptor scheme contributed the first two most significant features (LAT and HD) for instrument classification, however, as indicated by feature analysis, MFCC and perception-based features dominated in the ranked selections as well as SVM-based selections. It was also demonstrated that among the individual feature schemes it is the MFCC feature scheme that gave the best classification performance.

It is interesting to see that the feature schemes adopted in current research works are all highly redundant as assessed by the dimension reduction techniques. This may imply that an optimal and compact feature scheme remains to be found, allowing classifiers to be built fast and accurately. The finding of an embedding dimension as low as 4 or 5, however, may relate to the specific sound source files we used in this study and its scalability needs further verification.

On the other hand, in the classification of individual instruments, even the full feature set would not help much especially in distinguishing woodwind instruments. In fact, it is found in our experiments on solo passage classification that some MPEG-7 features are not reliable for giving robust classification results with the current fixed segmentation of solo passages. For instance, attack time is not selected in the feature scheme, but it can become a very effective attribute with the help of onset detection. All these indicate more research work in feature extraction and analysis is still necessary.

Apart from the timbral feature schemes we examined, there are other audio descriptors in the MPEG-7 framework that may contribute to better instrument classification, e.g. those obtained from global spectral analysis such as spectral envelop and spectral flatness [15]. Despite some possible redundancy with the introduction of new features, it would be interesting to investigate the possible gains that can be obtained through more study on feature analysis and selection, and how the proposed approach scales with increased feature numbers and increased amount of music samples.

In the future, we intend to investigate the feature extraction issue with the use of real world live recorded music data, develop new feature schemes, as well as experiment on finding better mechanisms to combine the feature schemes and improve the classification performance for more solo instruments.

References

[1] Y.-H. Tseng, "Content-based retrieval for music collections," in *Proc. of the 22nd ACM SIGIR International Conference on Research and Development in Information Retrieval*, 1999, pp. 176–182.

[2] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proceedings of IEEE*, vol. 92, no. 4, pp. 712–729, 2004.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, pp. 293–302, 2002.

[4] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proceedings of the 6th Inter. Conf. on Music Information Retrieval*, 2005, pp. 34–41.

[5] J. Marques and P. Moreno, "A study of musical instrument classification using gaussian mixture models and support vector machines," Compaq Computer Corporation, Tech. Rep. CRL 99/4, 1999.

[6] A. Eronen, "Comparison of features for music instrument recognition," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 19–22.

[7] G. Agostini, M. Longari, and E. Poolastri, "Musical instrument timbres classification with spectral features," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, 2003.

[8] S. Essid, G. Richard, and B. David, "Efficient musical instrument recognition on solo performance music using basic features," in *Proceedings of the Audio Engineering Society 25th International Conference*, no. 2-5. Audio Engineering Society, 2004, accessed 22.11.2005. [Online]. Available: http://www.tsi.enst.fr/%7Eessid/pub/aes25.pdf

[9] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, pp. 2–10, 1999.

[10] H. G. Kim, N. Moreau, and T. Sikora, "Audio classification based on MPEG-7 spectral basis representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 716–725, 2004.

[11] J. Aucouturier and F. Pachet, "Scaling up music playlist generation," in *Proc. IEEE International Conference on Multimedia and Expo*, vol. 1, 2002, pp. 105 – 108.

[12] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang, "Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification," in *Proc. of IEEE International Conference on Multimedia and Expo*, vol. 3, 2003, pp. 397–400.

[13] L. Ma, B. Milner, and D. Smith, "Accoustic environment classification," *ACM Transactions on Speech and Language Processing*, vol. 3, no. 2, pp. 1–22, 2006.

[14] A. Divakaran, R. Regunathan, Z. Xiong, and M. Casey, "Procedure for audio-assisted browsing of news video using generalized sound recognition," in *Proceedings of SPIE*, vol. 5021, 2003, pp. 160–166.

[15] ISO/IEC Working Group, "MPEG-7 overview," URL http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm, 2004, accessed 8.2.2007.

[16] A. T. Lindsay and J. Herre, "MPEG-7 audio - an overview," *J. Audio Eng. Soc.*, vol. 49, no. 7/8, pp. 589–594, 2001.

[17] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7," in *Proc. of International Computer Music Conference*, 2000, pp. 166–169.

[18] J. C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1064–1072, 2001.

[19] A. A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Proceedings of the 7th International Conference on Digital Audio Effects*, 2004, pp. 222–226.

[20] E. Benetos, M. Kotti, and C. Kotropoulos, "Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection," in *Proceedings of ICASSP 2006*, vol. V, 2006, pp. 221–224.

[21] I. Kaminskyj and T. Czaszejko, "Automatic recognition of isolated monophonic musical instrument sounds using *k*nnc," *Journal of Intelligent Information Systems*, vol. 24, no. 2/3, pp. 199–221, 2005.

[22] J. Eggink and G. J. Brown, "Instrument recognition in accompanied sonatas and concertos," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. IV, 2004, pp. 217–220.

[23] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[24] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[25] M. Grimaldi, P. Cunningham, and A. Kokaram, "An evaluation of alternative feature selection strategies and ensemble techniques of classifying music," School of Computer Science and Informatics, Trinity College Dublin, Tech. Rep. TCD-CS-2003-21, 2003.

[26] J. Qinlan, C4.5*: Programs for machine learning*. Morgan Kaufmann, 1993.

[27] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical recipes in C*. Cambridge University Press, 1988.

[28] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[29] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[30] C. Atkeson, A. Moore, and S. Schaal, "Locally weighted learning," *Artificial Intelligence Review*, vol. 11, pp. 11–73, 1997.

[31] IPEM, "IPEM-toolbox," URL http://www.ipem.ugent.be/Toolbox, accessed 10/9/2005.

[32] M. Slaney, "Auditory-toolbox," 1998, accessed 22.2.2007. [Online]. Available: http://rvl4.ecn.purdue.edu/ malcolm/interval/1998-010

[33] M. Casey, "MPEG-7 sound-recognition tools," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 737–747, 2001.

[34] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, 2005.

[35] K. D. Martin and Y. E. Kim, "Musical instrument identification: a pattern-recognition approach," *Journal of the Acoustical Society of America*, vol. 103, no. 3 pt 2, p. 1768, 1998.