Feature Analysis and Classification of Classical Musical Instruments: An Empirical Study

Christian Simmermacher, Da Deng, Stephen Cranefield

Department of Information Science, University of Otago, New Zealand
{ddeng,scranefield}@infoscience.otago.ac.nz

Abstract. We present an empirical study on classical music instrument classification. A methodology with feature extraction and evaluation is proposed and assessed with a number of experiments, whose final stage is to detect instruments in solo passages. In feature selection it is found that similar but different rankings for individual tone classification and solo passage instrument recognition are reported. Based on the feature selection results, excerpts from concerto and sonata files are processed, so as to detect and distinguish four major instruments in solo passages: trumpet, flute, violin, and piano. Nineteen features selected from the Mel-frequency cepstral coefficients (MFCC) and the MPEG-7 audio descriptors achieve a recognition rate of around 94% by the best classifier assessed by cross validation.

1 Introduction

Research in music data retrieval for commercial or non-commercial applications has been very popular in the last few years. Even though speech processing applications are well established, the growing use and distribution of multimedia content via the Internet, especially music, imposes some considerable technical challenges and demands more powerful musical signal analysis tools. New methods are being investigated so as to achieve semantic interpretation of low-level features extracted using audio signal processing methods. For example, a framework of low-level and high-level features given by the MPEG-7 multimedia description standard [1] can be used to create application specific description schemes. These can then be utilised to annotate music with a minimum of human supervision for the purpose of music analysis and retrieval.

There are many potential applications to be found for instrument detection techniques. For instance, detecting and analysing solo passages can lead to more knowledge about different styles of musical artists and can be further processed to provide a basis for lectures in musicology. Also various applications for audio editing, audio and video retrieval or transcription can be supported. Other applications include music genre classification [2], play list generation [3], and using audio feature extraction to support video scene analysis and annotation [4]. An overview of audio information retrieval and relevant techniques can be found in [5].

With this work we intend to eventually recognise classical instruments in solo musical passages with accompaniment, using features based on human perception, cepstral features, and the MPEG-7 audio descriptors. We try to find synergies and differences between these feature schemes so as to build a robust classification system. The performance of the feature schemes is assessed individually and in combination with each other.

This rest of the paper is organised as follows. Section 2 highlights a few recent relevant works 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, classification algorithms used, as well as our experiment procedures and settings. Empirical results based on the proposed approach are then presented in Section 4, followed by a discussion. Finally, we conclude the paper in Section 5.

2 Related Work

Various feature schemes have been proposed and adopted in the literature, and different computational models or classification algorithms have been employed for the purpose 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 (e.g. [6–8]). Marques and Moreno [6] built a classifier that can distinguish between eight instruments with a 70% accuracy rate using Support Vector Machines (SVM). Eronen [7] assessed the performance of MFCC features and spectral and temporal features such as amplitude envelope and spectral centroid etc. for instrument classification. He conducted Karhunen-Loeve Transform to decorrelate the features and then used k-nearest neighbours (k-NN) classifiers whose performance was then assessed using cross validation. The results favoured MFCC features, and violin and guitar were among the most poorly recognised instruments.

The MPEG-7 audio framework targets on the standardisation of the extraction and description of audio features [1]. The sound description of MPEG-7 audio features was assessed in [9] 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. [10] compared MFCC and MPEG-7 audio features for the purpose of sports audio classification, adopting hidden Markov models and a number of classifiers such as k-NN, Gaussian mixture models (GMM), AdaBoost, and SVM.

Brown and Houix [11] conducted a study on identifying four instruments of the woodwind family. Features used were cepstral coefficients, constant Q transform (CQT), spectral centroid, autocorrelation coefficients (AC), and time features. For classification a k-Means based GMM was used. Recognition success of the feature sets varied from 75%-85%.

Essid et al. [8] processed and analysed 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 derivates of MFCC were used as features. SVM yielded an average result of 76% for 35 features. A subsequent work from the same authors [12] used the same experimental setup but employed different features including AC and CQT, as well as amplitude modulated features. A feature selection technique was presented and features were classified pairwise with an expectation-maximisation based GMM. Best average results showed an accuracy of around 80%.

In [13], spectral features were extracted while the classification performance was assessed using SVM, k-NN, canonical discriminant analysis, and quadratic discriminant analysis, with the first and last being the best.

Livshin and Rodet [14] 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 large 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.

Eggink and Brown [15] presented a study on the recognition of five instruments (flute, oboe, violin and cello) in accompanied sonatas and concertos. GMM classifiers were employed on features reduced by a principal component analysis. The classification performance on a variety of data resources ranges from 75% to 94%, while mis-classification occurs mostly on flute and oboe (as violin).

In terms of feature analysis, some generic methods such as information gain (IG) and symmetric uncertainty (SU) were discussed in [16]. Grimaldi et al. [17] evaluated selection strategies such as IG and gain ratio (GR) for music genre classification. Some wavelet packet transform features, beat histogram features, and spectral features were extracted, selected, and classified by k-NN classifiers.

On the other hand, there are very limited resources available for benchmarking, so direct comparison of these various approaches would be hardly possible. Most studies have used recordings digitised from personal or institutional CD collections. McGill University Master Samples (MUMS) have been used in [13, 15], while the Iowa music samples were used in [7, 15].

3 Methodology

In this section, we present our computational approach for instrument classification. We will briefly introduce some common feature extraction schemes, the feature selection methods used, and the classification models. Our experiment model is then introduced, including data sources used, experiment procedures and resources made use of.

3.1 Feature Extraction

One of our main intentions is to investigate the performance of different feature schemes and find an optimal feature combination for robust instrument classification. 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.

The perception-based approach represents the instrument sound samples in a physiological way by calculating a nerve image. Three main steps are involved: simulation of the filtering of the outer and middle ear, simulation of the basilar membrane resonance in the inner ear, and simulation of a hair cell model. A second-order low-pass filter is applied for the outer and inner ear filtering. It has a 4 kHz resonance frequency that approximately simulates the overall frequency response of the ear. The basilar membrane is implemented via arrays of band-pass filters. They are divided into 40 channels with frequencies from 141 to 8877 Hz. Finally, the hair cell model uses half-wave rectification and dynamic range compression to act like an amplifier.

Among the temporal features, *zero-crossing rate* (ZCR) is an indicator for the noisiness of the signal and is normally found in speech processing applications; the *root-mean-square* (RMS) feature summarises the energy distribution in each frame and channel over time; the *spectral centroid* measures the average frequency weighted by amplitude of a spectrum; *bandwidth* shows a signal's frequency range by calculating the weighted difference in a spectrum; *flux* represents the amount of local spectral change, calculated as the squared difference between the normalized magnitudes of consecutive spectral distributions.

In our approach we first use the Harmonic Instrument Timbre Description Scheme of the MPEG-7 audio framework, which consists of seven feature descriptors: Harmonic Centroid (HC), Harmonic Deviation (HD), Harmonic Spread (HS), Harmonic Variation (HV), Log-Attack-Time (LAT), Temporal Centroid (TC) and Spectral Centroid (SC). This is only a subset of the eighteen descriptors provided by the MPEG-7 audio framework.

To obtain MFCC features, a signal needs to be transformed from frequency (Hertz) scale to mel scale and a discrete cosine transform converts the filter outputs to MFCC. Here, the mean (denoted as MFCCnM) and standard deviation (as MFCCnD) of the first thirteen linear values are extracted for classification.

Table 1 lists the 44 extracted features. The first 11 features are perception-based, the next 7 are MPEG-7 feature descriptors, and the last 26 are MFCC features.

3.2 Feature Selection

Feature selection techniques are often applied to optimise 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 to save computational time. However, care has to be taken that not too many features are removed. The effect of multiple features substituting each other could be desirable, since it is not exactly clear how musical timbre is described best.

To evaluate the quality of a feature 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 to other relevant features [18]. Eventually

Table 1. Feature Description

Feature No.	Description	Scheme
1	Zero Crossings	Perception-based
2-3	Mean and standard deviation of ZCR	
4-5	Mean and standard deviation of RMS	
6-7	Mean and standard deviation of Centroid	
8-9	Mean and standard deviation of Bandwidth	
10-11	Mean and standard deviation of Flux	
12	Harmonic Centroid Descriptor	MPEG-7 Timbre Description
13	Harmonic Deviation Descriptor	
14	Harmonic Spread Descriptor	
15	Harmonic Variation Descriptor	
16	Spectral Centroid Descriptor	
17	Temporal Centroid Descriptor	
18	Log-Attack-Time Descriptor	
19-44	Mean and standard deviation of the first 13 linear MFCCs	MFCC

it boils down to the modeling of correlation between two variables or features. Based on information theory, a number of indicators can be developed.

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

$$H(X) = -\sum_i P(x_i) log_2 P(x_i), \tag{1}$$

where $P(x_i)$ is the prior probabilities for all values of X. The entropy of X after observing another variable Y is then defined as

$$H(X|Y) = -\sum_j P(y_j) \sum_i (P(x_i|y_j) log_2 P(x_i|y_j)), \tag{2}$$

The Information Gain (IG) [19], indicating the amount of additional information about X provided by Y, is given as

$$\text{IG}(X|Y) = H(X) - H(X|Y) \tag{3}$$

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

The gain ration method (GR) normalised IG with an entropy item:

$$\text{GR}(X|Y) = \frac{\text{IG}(X|C)}{H(Y)} \tag{4}$$

A better symmetrical measure is defined as the *symmetrical uncertainty* [20][18]:

$$\text{SU} = 2\frac{\text{IG}(X|Y)}{H(X) + H(Y)} \tag{5}$$

To calculate these feature selection indexes, the feature sets need to be discretized beforehand.

3.3 Classification

The following classification algorithms are used in this study: condensed k-NN, which is a lazy learning method with an edited set of prototypes [21]; multilayer perceptron (MLP), which is a feedforward neural network using error back-propagation for training; and support vector machine, which is a statistical learning algorithm and has been implemented in a number of machine learning toolboxes.

3.4 Experiment settings

In this study we tackle the music instrument classification problem in two stages:

- Instrument tone classification using samples of individual instruments.
- Solo instrument detection and classification.

For these experiments all audio features are extracted using the IPEM Toolbox [22] and Auditory Toolbox [23], and an implementation of the MPEG-7 audio descriptors by Casey [24] is used. Weka [25] is used for feature selection, and for classification using SVM and MLP. The condensed k-NN algorithm is implemented separately in Java.

Single instrument classification Samples used in the first experiment are taken from the Iowa Music Samples 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 recorded in the same acoustic environment (anechoic chamber) under the same conditions. We realise that this is a strong constraint and our result may not generalise to a complicated setting such as dealing with live recordings of an orchestra. The purpose of this experiment, however, is to test the behaviour of the feature schemes, evaluate the features using feature selection, and test the performance of different classifiers. It is also important for us to use some benchmark data also used in other research for this purpose.

Solo instrument classification For the second experiment, instrument classification is to perform on solo samples. These sample phrases are often polyphonic, therefore more challenging than the first experiment. One representative instrument of each class is chosen. The instruments are: trumpet, flute, violin, and piano. To detect the right instrument in solo passages, a classifier is trained on short monophonic phrases. Ten-second long solo excerpts from CD recordings are tested on this classifier. The problem here is that the test samples are 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 be the most dominant and the features derived from the harmonic spectrum are assumed to work. In order to get a smaller but more robust feature scheme, a feature selection algorithm is applied.

The samples for the four instruments are taken from CD recordings from private collections and the University of Otago Library. Each instrument has at least five sources and each source is taken either for training or testing to guarantee the independence of the data set. As seen in Table 2, three sources are used for the training set and at least nine minutes of two second monophonic phrases are extracted from them. The test set has two sources for trumpet and flute, and three sources for piano and violin. Passages of around ten-second length are segmented into two second phrases with 50% overlap. The difference in the number of test samples is due to this process.

Table 2. Data Sources used in solo instrument classification

Sources	Training set		Test set	
Trumpet (5)	9 min	270 samples	3.3 min	181 samples
Piano (6)	10.6 min	320 samples	4 min	219 samples
Violin (6)	10 min	300 samples	4 min	215 samples
Flute (5)	9 min	270 samples	3.3 min	185 samples
Total (22)	38.6 min	1160 samples	14.6 min	800 samples

The test set includes different recordings of the four instruments. Samples of the piano are pure solo passages. 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 solos and sometimes with flute and string accompaniment.

4 Results

4.1 Instrument tone classification

Feature selection For this purpose, we first simply the instrument classification problem by grouping the instruments into four major classes: piano, brass, string and woodwind. For this 4-class task, the best 20 features of the three selection methods are shown in Table 3. 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 and represent the data collectively. It is necessary to mention that the standard deviation of the MFCC is predominantly present in all three selections. Also the measures of the centroid and bandwidth, as well as one representative of flux, zero crossings and energy can be found in each of them.

Table 3. Feature selection for single tones

Rank	IG		GR		SU	
	Relevance	Feature	Relevance	Feature	Relevance	Feature
1	0.8154	LAT	0.531	LAT	0.4613	LAT
2	0.6153	HD	0.527	HD	0.3884	HD
3	0.419	FluxD	0.323	MFCC2M	0.2267	BandwidthM
4	0.3945	BandwidthM	0.297	MFCC12D	0.219	FluxD
5	0.3903	MFCC1D	0.27	MFCC4D	0.2153	RMSM
6	0.381	MFCC3D	0.266	BandwidthM	0.2084	MFCC1D
7	0.3637	RMSM	0.264	RMSM	0.1924	MFCC4M
8	0.3503	BandwidthD	0.258	MFCC13D	0.1893	MFCC11D
9	0.342	MFCC4M	0.245	MFCC2D	0.1864	MFCC3D
10	0.3125	MFCC11D	0.24	MFCC11D	0.1799	BandwidthD
11	0.3109	ZCRD	0.235	MFCC7D	0.1784	MFCC2M
12	0.2744	CentroidD	0.229	FluxD	0.1756	MFCC4D
13	0.2734	MFCC8D	0.224	MFCC1D	0.171	MFCC7D
14	0.2702	MFCC6D	0.22	MFCC4M	0.1699	MFCC12D
15	0.2688	MFCC7D	0.215	CentroidM	0.1697	ZCRD
16	0.2675	ZC	0.211	SC	0.1653	CentroidD
17	0.2604	MFCC4D	0.209	MFCC5M	0.161	CentroidM
18	0.2578	CentroidM	0.208	CentroidD	0.1567	MFCC13D
19	0.2568	MFCC10M	0.195	HC	0.1563	SC
20	0.2519	MFCC10D	0.191	MFCC1M	0.1532	MFCC8D

Choice of classifier scheme Next, we work onto examining the choice of feature sets together with the classification algorithms, so as to determine a final classification scheme. Three data analysis methods (k-NN, SVM, MLP) are compared for this classification task. Each of them splits 66% of the data into training instances and takes the rest for testing. The percentage split takes the distribution of class values into account so that each class is reasonably well represented in both training and testing sets. In a first step a classifier is trained on all features. The first 30, 20 and ten features from the information gain filter are taken as the reduced feature set, since they show better results than gain ratio and symmetrical uncertainty. The results are given in Table 4.

Table 4. Performance of three classifiers for the four classes.

Feature Scheme	k-NN	SVM	MLP
All 44 features	65.15%	82.58%	92.65%
Best 30	63.64%	82.58%	91.91%
Best 20	58.33%	75%	91.91%
Best 10	56.06%	59.09%	88.97%

Table 5. Performance of the feature sets in classifying the 4 classes (10 CV)

Feature set	Piano	Brass	String	Woodwind	Average
MFCC (26)	99%	92%	87%	64%	85.5%
MPEG-7 (7)	99%	67%	57%	48%	67.75%
IPEM (11)	100%	76%	85%	36%	74.25%
MFCC-MPEG-7 (33)	100%	92%	95%	71%	89.5%
MFCC-IPEM (37)	98%	93%	92%	78%	90.25%
MPEG-7-IPEM (18)	99%	82%	93%	48%	80.5%
All (44)	100%	90%	97%	73%	90%

The k-NN classifier achieved its best performance with three nearest neighbours. For all features and the best 30 features 200 prototypes are found, the best 20 have 217 prototypes and the best ten have 227. The SVMs use a polynomial kernel with an exponential of 9 and the C value is set to 10. Ninety-two support vectors are used for all features and the best 30, 133 for 20 features, and 235 for ten features. The MLP is trained over 500 epochs with a learning rate of 0.3 and a momentum of 0.2. The accuracy increases with the amount of features in all three classifications. The variance of the results in MLP is not as large as in the other two, and it also shows the highest recognition rate.

The performance of different feature set combinations are then assessed with the best classifier - MLP. A 10-fold cross validation process is employed to obtain the results as given in Table 5.

In terms of average performance, the MFCC-IPEM set shows the closest results compared to all 44 features. The 18 features from the MPEG-7-IPEM set have lowest combination result. Generally, the sum of features shows better results. However, between 33, 37 and 44 features there is not even one percent difference. MFCCs are included in all these, being probably the most significant features.

The piano could be classified by all feature sets near to perfect. The MPEG-7 and IPEM sets have problems identifying brass instruments, only the IPEM set could increase the performance of MFCC for this task. String instruments have a high recognition rate except for the MPEG-7 feature set. But combined with MFCC the rate improves to 95%, which is good considering the amount of features (33). All individual feature sets had problems classifying the woodwind class, which is probably because of the few samples in relation to the number of instruments. Only the combination of MFCC-IPEM upgraded the performance to a maximum of 78%. This capability of MFCC-IPEM makes it the best working combination on average for all four instrument classes.

Instrument Classification Based on the result given above, MLP is chosen as the classifier for further experiments, in which all 20 instruments are directly differentiated against each other. The Iowa samples are used to train a classifier with all 44 features. The confusion matrix of the 20-instrument classification is given in Table 6, where the 10-fold cross validation results are shown. At the

bottom it also shows the combined classification rate for the four instrument groups, with 'piano' being the best, and 'woodwind' the worst.

Table 6. Confusion matrix for all 20 instruments with 10-fold CV using all 44 features.

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	100																			
b=tuba		20																		
c=trumpet			19			1														
d=horn	1			19																
e=tenoTrombone	1		1		18															
f=baseTrombone					6	14														
g=violin							25													
h=viola	1						1	24												
i=bass	1	1							23											
j=cello										25										
k=sax											8	1							1	
l=altoSax												6			1		1		2	
m=oboe	1									1			7						1	
n=bassoon														10						
o=flute															9			1		
p=altoFlute		1														8	1			
q=bassFlute																2	8			
r=bassClarinet	1								2	1		1			1			5		
s=bbClarinet											1				2			1	6	
t=ebClarinet												1	1			1			1	6
Combined	100%	90%					97%				73%									

4.2 Solo instrument detection

Feature selection Again we apply the three feature selection measures for the training features. The result is shown in Table 7. All selection techniques indicate the same features (except MFCC6M and CentroidD) and also their ranking is nearly similar. It is to notice that nearly all IPEM features are represented (except CentroidD in information gain and symmetrical uncertainty), as well as the means of the first seven MFCC. For the MPEG-7 scheme SC, HC, HS, and HV work best.

Again, the three feature selection filters extract similar groups of features. It seems that among the 44 features, log-attack time, energy features and all standard deviations of the MFCCs are not or only minimal relevant. It is not surprising that LAT is not relevant, since the phrases are cut sequentially at two second intervals, thus there is no proper information of the instrument attack phase. Even if this information would be present, it could be horizontally

Table 7. Feature selection for solo passages

Rank	IG		GR		SU	
	Relevance	Feature	Relevance	Feature	Relevance	Feature
1	1.0028	SC	0.4653	MFCC2M	0.4819	MFCC2M
2	0.99748	MFCC2M	0.4413	SC	0.4699	SC
3	0.97115	HC	0.401	HC	0.4396	HC
4	0.82191	ZCRM	0.3477	ZC	0.3712	ZCRM
5	0.78518	ZC	0.338	ZCRM	0.3691	ZC
6	0.72037	MFCC3M	0.2808	HD	0.309	MFCC3M
7	0.62972	CentroidM	0.2702	MFCC3M	0.2954	HD
8	0.62191	HD	0.2631	CentroidM	0.2869	CentroidM
9	0.52701	ZCRD	0.2475	ZCRD	0.2555	ZCRD
10	0.51799	HS	0.247	HS	0.2531	HS
11	0.50303	MFCC4M	0.2337	MFCC1M	0.238	MFCC4M
12	0.43957	MFCC1M	0.231	MFCC10M	0.2268	MFCC1M
13	0.41417	MFCC10M	0.2255	MFCC4M	0.2186	MFCC10M
14	0.37883	FluxM	0.1793	FluxM	0.1844	FluxM
15	0.3643	MFCC5M	0.1752	MFCC7M	0.1752	MFCC7M
16	0.34954	MFCC7M	0.1573	MFCC5M	0.1689	MFCC5M
17	0.30444	BandwidthM	0.1517	BandwidthM	0.1521	BandwidthM
18	0.28482	FluxD	0.147	FluxD	0.1448	FluxD
19	0.22816	MFCC6M	0.1386	CentroidD	0.1175	BandwidthD
20	0.22358	BandwidthD	0.1235	BandwidthD	0.1147	MFCC6M

masked by successive tones or other instruments. Apart from LAT, the standard deviation of the MFCCs as well as TD (MPEG-7) are discarded.

The dynamics of the phrase could also be the cause for the decline in relevance of energy features. Phrases are not static like the instrument tones in the first experiment. Composition and playing style may cause the instruments' dynamic ranges difficult to extract. Successive tones and other instruments can also mask the instrument vertically.

For this task the original feature set is reduced to only 28 features. It consists of the means of the MFCC, the MPEG-7 features without LAT, and the IPEM features without RMS. The new set of features was calculated from the 1160 training and 800 test samples. This accounts for a 59% split of the data set for training.

The evaluation of the feature set using a MLP to classify the four instruments is shown in Table 8. The MFCC set alone has a high recognition rate with only 13 features. They sometimes mistake trumpet for piano, but some more errors exist in the representation of piano; it is sometimes misinterpreted as violin or flute.

All 28 features achieved a 93.5% recognition rate. Furthermore, the sets of the selected best 20 and 10 features had a classification rate of 92.13% and 82.88% respectively. We can see that the higher the number of features the higher the increase in accuracy.

Table 8. Performance on feature sets in combination for four instruments

MLP	Trumpet	Piano	Violin	Flute	Average
MFCC (13)	93.4%	76.3%	97.7%	100%	91.38%
MPEG (6)	40.9%	69.9%	91.6%	82.2%	72%
IPEM (9)	48.6%	74.9%	93.5%	91.9%	77.9%
MFCC-MPEG-7 (19)	91.7%	88.6%	97.7%	99.5%	94.25%
MFCC-IPEM (22)	93.9%	74%	100%	99.5%	91.4%
MPEG-7-IPEM (15)	83.4%	68%	96.7%	94.1%	85.25%
All 28 features	95%	82.6%	98.1%	99.5%	93.5%

The highest average accuracy for detecting the four instruments is 94.25%, achieved by combining the MPEG-7 features with MFCCs. Interestingly, this feature set of only 19 features performs better than the selected best 20 features, indicating that feature selection may not guarantee the exact best performance.

A change to shorter one second segments for training and testing shows similar results but with a tendency of reducing to lower recognition rates.

Piano is often misclassified as flute and trumpet is confused with violin using IPEM and MPEG-7 features alone. However, these two sets have less than ten features and probably do not capture all necessary information from the signal. Violin and flute are classified excellently by all feature sets. The IPEM and MPEG-7 feature sets have a low recognition rate for the trumpet samples. In combination the accuracy can be boosted to over 90%. The piano is the hardest instrument to classify, even though its training and test samples are mostly pure solo passages without accompaniment. The 88.6% accuracy of the MFCC-MPEG-7 set for piano makes it the best working combination on average.

The confusion matrix for the four instruments (Table 9) shows that even with all 28 features employed the classifier cannot fully distinguish between piano and violin, and piano and flute. Furthermore, sometimes trumpet and violin are misinterpreted as piano.

Table 9. Confusion matrix for four instruments classification with 28 features

Instruments	Predicted As			
	Trumpet	Piano	Violin	Flute
Trumpet	172	9	0	0
Piano	1	181	7	30
Violin	0	4	211	0
Flute	0	0	1	184

4.3 Discussion

The feature selection for the two classification problems given above produces similar results, agreeing to that MFCCs are the most important features for

instrument classification. There are some interesting difference, however. In the solo passage experiments, e.g., the standard deviation features of MFCC are found to be irrelevant. Rather, the mean values of the MFCC are the most robust feature set.

In general, the solo instrument detection classification is a more challenging problem, dealing with mostly polyphonic samples. The highest recognition rate is achieved by a combination with the MPEG-7 set. Spectral Centroid and Harmonic Centroid from the MPEG-7 scheme have a high relevance in the feature selection and perhaps could capture more information in combination with the MFCC scheme. The combination MFCC-IPEM was found to be unable to improve the result achieved by MFCC alone.

Flute and violin are the instruments with the highest classification accuracy. This is different from the findings in [7] and [15]. The IPEM and MPEG-7 features sets have problems representing the trumpet. Better results are achieved with the MFCC set or generally with a combination of the feature sets. Throughout piano is detected with the lowest average accuracy of around 75%, contrary to the finding in the experiments on the samples of single instruments.

Generally, the high classification rates are possibly due to the distinctive acoustic properties of the instruments, since they originate from different families. It is not claimed that these results generalise. The accuracy is likely to decrease when more instruments and more sources of samples are introduced.

Using feature selection filters, the most informative features for solo passages are found to be the spectral and harmonic centroids, zero-crossing rate, and generally the first seven MFCCs. It is not surprising that, for instance, time-dependent features cannot represent the signal, since a strict sequential segmentation is applied, leaving incomplete spatial cues, and destroying the context of the temporal information.

Another possibility can be to employ segmentation at onset time, so that it make sense to apply time-dependent features again. However, the detection of onset is itself a challenging problem. Nevertheless, using spectral features alone can achieve a good performance as shown in our experiments, where a simple sequential segmentation of the solo passage was implemented.

5 Conclusion

In this paper, we studied feature extraction and evaluation for the problem of instrument classification. The main contribution is that we bring three major feature extraction schemes under investigation, have them analysed using feature selection measures based on information theory, and their performance assessed using classifiers undergoing cross validation. In the first experiment of instrument tone classification, a publicly available data set is used, allowing for the possibility of benchmark comparison. For instance, Iowa music samples were also used in [7], but our results on instrument family classification and instrument tone classification are much higher on almost all common instruments in both studies.

Three feature selection measures are adopted and produce similar feature selection outputs, which basically aligns with the performance obtained through classifiers, especially MLP. We note that the use of an MLP is rather uncommon for recognising musical tones and phrases as shown from the literature. It however produces favourable results both on tone and solo passage classification. This finding may not generalise, but we will assess its performance using more music samples, and compare the performance of more classification models. Also, by conducting linear and non-linear principal component analyses, their dimension reduction and de-noising effects may also enhance the final classification outcome.

We have covered only four instruments in the solo passage experiments. The intention is to distinguish major classical solo instruments in accompanied solo passages in concertos or sonatas. There are few works concentrating on polyphonic music without separating the signal into its sound or instrument sources. Detecting the range of orchestral instruments therein still needs a considerable effort of research. A comparison study on analysing musical passages is hard to achieve, as till now there is no free accessible common data set which could be used as a basis for further work.

In the future, we intend to investigate the feature extraction issue with more real-world music data, develop new feature schemes, as well as experiment on finding better mechanisms to combine the feature schemes and the classification models in order to achiever better performance on more solo instruments.

References

1. ISO/IEC Working Group: MPEG-7 overview. URL http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm (2004) Accessed 8.2.2006.
2. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on speech and audio processing **10** (2002) 293–302
3. Aucouturier, J., Pachet, F.: Scaling up music playlist generation. In: Proc. ICME. Volume 1. (2002) 105 – 108
4. Divakaran, A., Regunathan, R., Xiong, Z., Casey, M.: Procedure for audio-assisted browsing of news video using generalized sound recognition. In: Proc. SPIE. Volume 5021. (2003) 160–166
5. Foote, J.: An overview of audio information retrieval. Multimedia Systems **7** (1999) 2–10
6. Marques, J., Moreno, P.: A study of musical instrument classification using gaussian mixture models and support vector machines. Technical Report CRL 99/4, Compaq Computer Corporation (1999)
7. Eronen, A.: Comparison of features for music instrument recognition. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. (2001) 19–22
8. Essid, S., Richard, G., David, B.: Efficient musical instrument recognition on solo performance music using basic features. In: Proceedings of the Audio Engineering Society 25th International Conference. (2004) Accessed 22.11.2005.
9. Peeters, G., McAdams, S., Herrera, P.: Instrument sound description in the context of mpeg-7. In: Proc. of Inter. Computer Music Conf. (2000) 166–169

10. Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, T.: Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification. In: Proc. of ICME. Volume 3. (2003) 397–400
11. Brown, J.C., Houix, O.: Feature dependence in the automatic identification of musical woodwind instruments. Journal of the Acoustical Society of America **109**(3) (2001) 1064–1072
12. Essid, S., Richard, G., David, B.: Musical instrument recognition by pairwise classification strategies. IEEE Trans. on Speech and Audio Processing (to appear) (2006) Accessed 2.12.2005.
13. Agostini, G., Longari, M., Poolastri, E.: Musical instrument timbres classification with spectral features. EURASIP Journal on Applied Signal Processing (1) (2003)
14. Livshin, A.A., Rodet, X.: Musical instrument identification in continuous recordings. In: Proceedings of the 7th International Conference on Digital Audio Effects. (2004)
15. Eggink, J., Brown, G.J.: Instrument recognition in accompanied sonatas and concertos. In: Proc. of ICASSP. Volume IV. (2004) 217–220
16. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Machine Learning Research **5** (2004) 1205–1224
17. Grimaldi, M., Cunningham, P., Kokaram, A.: An evaluation of alternative feature selection strategies and ensemble techniques of classifying music. In Mladenic, D., Paa, G., eds.: Workshop in Multimedia Discovery and Mining at ECML/PKDD-2003. (2003)
18. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proc. 20th Intl Conf. Machine Learning. (2003) 856–863
19. Qinlan, J.: *C4.5*: Programs for machine learning. Morgan Kaufmann (1993)
20. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical recipes in C. Cambridge University Press (1988)
21. Niemann, H.: Klassifikation von Mustern. Springer Verlag (1983)
22. IPEM: IPEM-toolbox. URL http://www.ipem.ugent.be/Toolbox (2005) Accessed 10.9.2005.
23. Slaney, M.: The Auditory Toolbox. Technical Report 1998-010, Interval Research Corporation (1998) URL http://rvl4.ecn.purdue.edu/∼malcolm/interval/1998-010. Accessed 22.11.2005.
24. Casey, M.: MPEG-7 sound-recognition tools. IEEE Trans. on Circuits and Systems for Video Technology **11**(6) (2001) 737–747
25. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Second edn. Morgan Kaufmann, San Francisco (2005)