A Study on Feature Analysis for Musical Instrument Classification

Article *in* IEEE TRANSACTIONS ON CYBERNETICS · May 2008

DOI: 10.1109/TSMCB.2007.913394 · Source: PubMed

3 authors, including:

Jeremiah D. Deng
University of Otago

Stephen Cranefield
University of Otago

A Study on Feature Analysis for Musical-Instrument Classification

Jeremiah D. Deng, *Member, IEEE*, Christian Simmermacher, and Stephen Cranefield

***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.**

***Index Terms*—Feature extraction, feature selection, music, pattern classification.**

I. 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]–[8]).

Instrument-detection techniques can have many potential applications. For instance, detecting and analyzing solo passages can lead to more knowledge about the different musical styles and can be further utilized to provide a basis for lectures in musicology. Various applications for audio editing and 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 for achieving semantic interpretation of low-level features extracted by audio-signal-processing methods. For example, a framework of low- 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 then generated, and their performance in instrument classification was further tested with a few classifiers using cross-validations. Three feature schemes were considered: features 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 to gain insight on the right dimensionality for feature selection. Our aim was to find the differences and synergies between the 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 II reviews the recent relevant work on musical-instrument recognition and audio-feature analysis. Section III outlines the approach that 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 IV. We summarize the findings and conclude the paper in Section V.

II. 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 and instrument classifications. Marques and Moreno [5] built a classifier that can distinguish between eight instruments with 70% accuracy using the support vector

Manuscript received February 13, 2007; revised July 31, 2007. This paper was recommended by Associate Editor P. S. Sastry.

J. D. Deng and S. Cranefield are with the Department of Information Science, University of Otago, Dunedin, New Zealand (e-mail: ddeng@infoscience.otago.ac.nz; scranefield@infoscience.otago.ac.nz).

C. Simmermacher is with IBM Germany, 71083 Herrenberg, Germany (e-mail: chris-sim@gmx.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2007.913394

machines (SVM). Eronen [6] assessed the performance of MFCC, 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 neighbor (k-NN) classifiers were used, with their performance assessed through cross-validation. The results favored the MFCC features, and violin and guitar were among the most poorly recognized instruments.

The MPEG-7 audio framework targets the 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 could be reliably applied in assessing the similarity of musical sounds. Xiong *et al.* [12] compared the MFCC and MPEG-7 audio features for the purpose of sports-audio classification, adopting the hidden Markov models (HMMs) 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 in 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, and autocorrelation coefficients. For classification, a scheme using the Bayes decision rules was adopted. The recognition rates based on the feature sets varied from 79% to 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 the 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.

Kostek [2] studied the classification of 12 instruments played under different articulations. She used multilayer neural networks trained on wavelet-transform features and MPEG-7 descriptors. 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.* [19] dealt with isolated monophonic instrument-sound recognition using k-NN classifiers. Features used included MFCC, constant-Q-transform spectrum frequency, root-mean-square (rms) amplitude envelope, spectral centroid, and multidimension-scaling (MDS) analysis trajectories. These features underwent principal component analysis (PCA) for reduction to a total dimensionality of 710. The 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 [20] 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%, whereas misclassification occurred mostly for flute and oboe (with both classified as violin). Essid *et al.* [8] processed and analyzed solo musical phrases from ten instruments. Each instrument was represented by 15 min of audio material from various CD recordings. Spectral features, audio-spectrum flatness, MFCC, and derivatives of MFCC were used as features. An SVM classifier yielded an average accuracy of 76% with 35 features. Livshin and Rodet [21] evaluated the use of monophonic phrases for instrument detection 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 real-time 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.* [22] adopted the branch-and-bound search to extract a six-feature subset from a set of MFCC, MPEG-7, and other audio spectral features. A nonnegative matrix factorization algorithm was used to develop the classifiers, gaining an accuracy of 95.2% for six instruments.

With the emergence of many audio-feature schemes, feature analysis and selection has been gaining more attention recently. A good introduction on feature selection was given in the work of Guyon and Elisseeff [23], outlining the methods of correlation modeling, 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 [21] 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%, which is 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; therefore, direct comparison of these various approaches is hardly possible. Most studies have used recordings digitized from personal or institutional CD collections. The McGill University Master Samples (http://www.music.mcgill.ca/resources/mums/html/mums.html) have been used in some studies [7], [19], [20], whereas the music samples from the UIOWA MIS Database http://theremin.music.uiowa.edu/) were also widely used [18], [20], [22].

III. Feature Analysis and Validation

A. *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,

TABLE I
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

oboe and clarinet belong to the woodwind category, horn and trumpet are brass instruments, and piano is usually classified as a percussion instrument. Sounds produced by these instruments bear different acoustic attributes. A few characteristics can be obtained from their 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.

B. *Feature Extraction for Instrument Classification*

Because of the complexity of modeling instrument timbre, various feature schemes have been proposed through acoustic study and pattern-recognition research. Our main intentions are to investigate the performance of different feature schemes and find a good feature combination for a robust instrument classifier. Here, we consider 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, whereas the last is based on spectral analysis. These features, 44 in total, are listed in Table I. Among them, the first 16 are perception-based features, the next 7 are MPEG-7 descriptors, and the last 26 are MFCC features.

1) Perception-Based Features: To extract perception-based features, music sound samples were segmented into 40-ms frames with 10-ms overlap. Each frame signal was analyzed by 40 bandpass filters centered at Bark-scale frequencies. The following are some important perceptual features used in this paper.

1) Zero-crossing rate (ZCR), an indicator for the noisiness of the signal, which is often used in speech-processing applications

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

where N is the number of digf samples in the frame, and F_n is the value of the nth sample of a frame.

2) Root mean square (rms), which summarizes the energy distribution in each frame

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

3) Spectral centroid, which measures the average frequency weighted by the 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 kth channel, $K = 40$ is the number of channels, and $P(f_k)$ is the spectrum amplitude on the kth channel.

4) Bandwidth (also referred to as the centroid width), which shows the frequency range of a signal weighted by its spectrum

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

5) Flux, representing the amount of local spectral change, which is 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 the mean value and standard deviation were used as the feature values for each sound sample.

2) MPEG-7 Timbral Features: Instruments usually have some unique properties that can be described by their harmonic spectra and their temporal and spectral envelopes. The MPEG-7 audio framework [15] endeavors to provide a complete feature set for the description of harmonic instrument sounds. We consider in this paper 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, whereas the last two belong to the timbral temporal scheme. Note that the SC feature value was obtained from the spectral analysis of the entire sample signal; thus, it is similar to but different from the CentroidM of the perception-based features. CentroidM was aggregated from the centroid feature extracted from short segments within a sound sample.

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

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

The mel scale has 40 filter channels. The first extracted filterbank output is a measure of power of the signal, and the following 12 linearly 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. The mean and standard deviation of the first 13 coefficients thus obtained were extracted for classification.

C. Feature Selection

Feature-selection techniques are often necessary for optimizing the feature sets used in classification. This way, redundant features are removed from the classification process, and the dimensionality of the feature set is reduced 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 modeling 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 prediscretized feature set, the "noisiness" of the feature X can be measured as the entropy, which is 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 ith discretized value of X. The entropy of X after observing another variable Y is then defined as

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

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

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

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

The GR method normalizes IG by an entropy term

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

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

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

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 aforementioned criteria [24], [25], 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.

D. Feature Analysis by Dimension Reduction

Standard dimension-reduction or MDS techniques such as PCA and Isomap [29] are often used to estimate an embedding dimension of the high-dimensional feature space. PCA projects high-dimensional data into low-dimensional space while preserving the maximum variance. It has been found rather effective in pattern-recognition tasks such as face 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 paper, we used PCA and Isomap to explore the sparseness of the feature space and examine the residuals of the chosen dimensionality to estimate how many features at least should be included in a subset. The performance of the selected subsets was then compared with that of the reduced and transformed feature space obtained by MDS.

E. Feature Validation via Classification

Feature-combination schemes generated from the selection rankings were then further assessed using classifiers under cross-validation. The following classification algorithms were used in this paper: k-NN, an instance-based classifier weighted by the reciprocal of distances [30]; naive Bayes, employing Bayesian models in the feature space; multilayer perceptron (MLP) and radial basis functions (RBFs), which are both neural

TABLE II
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

network classifiers; and SVM, which is a statistical learning algorithm and has been widely used in many classification tasks.

IV. EXPERIMENT

A. *Experiment Settings*

We tackled the musical-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 Waikato Environment for Knowledge Analysis (Weka) [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 (an anechoic chamber) under the same conditions. We realized that this was a strong constraint, and our result might 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 CD recordings from private collections and the University of Otago Library.

B. *Instrument-Family Classification*

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 generated by the three selection methods are shown in Table II. All of them indicate that LAT and HD are the most relevant features. 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 [21], where MPEG-7 descriptors were not considered. However, they also included bandwidth (spectral spread), MFCC, and SC.

Classifiers were then employed to assess the quality of feature selection. A number of algorithms, including naive Bayes, k-NN, MLP, RBF, and SVM, were compared on classification performance based on tenfold 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 minimal optimization, with RBF kernels and a complexity value of 100, with all attributes being standardized. Pairwise binary SVM classifiers were trained for this multiclass problem, with between 10 and 80 support vectors being created for each SVM. The structure of MLP was automatically defined in the Weka implementation, and each MLP was 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 IG 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, are not considered here. The performance of these reduced sets was compared with the original full set with all 44 features. The results are given in Table III.

These can be contrasted with the results presented in Table IV, 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,

TABLE III
CLASSIFIER PERFORMANCE (IN PERCENTAGE) 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 IV
PERFOMANCE (IN PERCENTAGE) 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

TABLE V
PERFORMANCE (IN PERCENTAGE) IN CLASSIFYING THE FOUR CLASSES (TENFOLD CROSS-VALIDATION)

AQ3

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

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 III. The performance of the "Selected 17" feature set is very close to that of the full feature set. The k-NN classifier performs even slightly better with the reduced feature set.

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 one-NN classifiers to assess the contribution from each individual feature scheme and improvements achieved through scheme combinations. Apart from combining the schemes one by one, another option was also considered: picking the top 50% ranked attributes from each feature scheme, resulting in a 21-dimension composite set, called the "Top 50% mix." The results are presented in Table V. Aside from 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 reveals that MPEG-7 features give better results than MFCC for the classification of sports-audio scenes such as applause, cheering, music, etc. The difference was however marginal (94.73% versus 94.60%). Given that the scope of this paper is much narrower, this should not be regarded as a contradiction. Indeed, some researchers also found more favorable results using MFCC instead of MPEG-7 for instrument classification [8], [10].

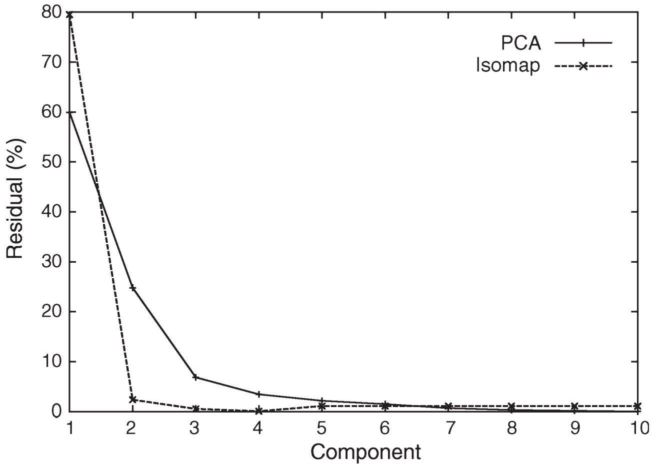

Fig. 1. Graphical representation of the reduced components. The x-axis gives the component number, and the y-axis gives the relevant normalized residual (in percentage). Only ten components are shown.

AQ4

In terms of average performance of combination schemes listed in Table V, the MFCC+MPEG-7 set produced the best results, whereas the MPEG-7+IPEM set with 18 features gave the poorest result. It is observed that the inclusion of MFCC is most beneficial to the woodwind and string families, whereas 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, the difference among 33, 37, and 44 features is almost negligible. It is interesting to note that the "Selected 17" feature set produced very good performance. The "Top 50% mix" set produced a performance as high as 93%, slightly worse than that of the "Best 20" set, probably due to the fact that the selection was 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 rather accurately classified on most feature sets. The MPEG-7 and IPEM sets seem to 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 despite their strong redundancy.

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 III 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 ten components using these methods are shown in Fig. 1. The 3-D projection of the Isomap algorithm, generated by selecting the first three coordinates from the resulting embedding, is shown in Fig. 2. The separability of the four classes already

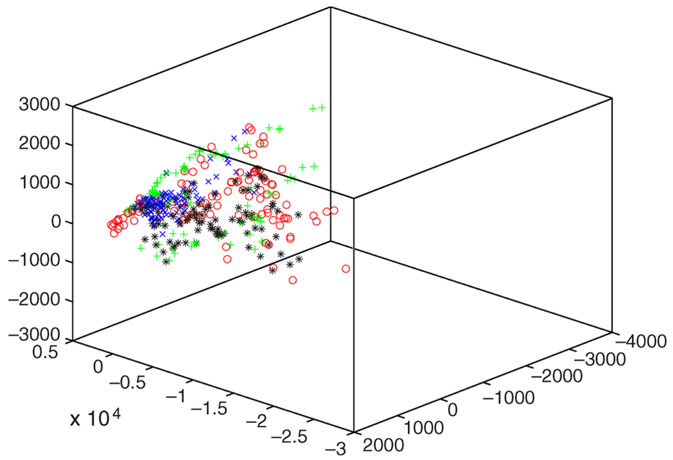

Fig. 2. Three-dimensional embedding of the feature space. There are 400 instrument samples, each with its category labeled: ×—"piano," ○—"string," +—"brass," and ∗—"woodwind." The three axes correspond to the transformed first three dimensions generated by Isomap.

starts to emerge with three dimensions. For both methods, the residual falls under 0.5% after the fourth component, although the dropping reported by Isomap is more significant. This suggests that the data manifold of the 44-D feature space may have an embedded dimension of four or five only.

As a test, the first five principal components (PCs) of the complete feature set were extracted, resulting in weighted combinations of MFCC, IPEM, and MPEG-7 features. A one-NN classifier trained with the five PCs reports an average accuracy of 88.0% in a tenfold cross-validation, very close to that of the "Best 5" selection given in Table III. This further confirms that there is strong redundancy within and between the three feature schemes.

C. *Instrument Classification*

1) Individual-Instrument Sound Recognition: Next, all 20 instruments were directly distinguished from each other. We chose to use one-NN classifiers as they worked very quickly and gave almost the same accuracies 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 VI. Instrument "a" is piano, and instruments "b–f" belong to the brass type, "g–j" to the string type, and "k–t" to the woodwind type.

The overall average classification accuracy is 86.9%. The performance, in general, is quite satisfactory, particularly for piano and string instruments. Only one out of 20 piano samples was wrongly classified (as oboe). Among the 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 into 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.

2) Instrument Recognition in Solo Phrases: Finally, a preliminary experiment on instrument recognition in solo phrases was conducted. For this experiment, one representative instrument of each instrument type was chosen. These were as follows: trumpet, flute, violin, and piano. To detect the right instrument in solo passages, a classifier was trained on short monophonic phrases. Solo excerpts from CD recordings were tested on this classifier. The problem here is that these solo phrases were recorded with accompaniment; thus, they were often polyphonic in nature. Selecting fewer and clearly distinguishable instruments for the trained classifier helps 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.

The samples for the four instruments were taken from live CD recordings. 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. Passages of around 10-s length were segmented into 2-s phrases with 50% overlap. Shorter segments seemed to have a tendency to lower classification rates. The amount of music samples was basically balanced across the four instrument types, as seen in Table VII.

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

AQ5

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

D. *Discussion*

The scopes of some current studies and performance achieved are listed in Table IX, where the number of instruments and the classification accuracies (in percentages) of instrument-family and individual-instrument classifications are listed. It can be seen that our results are better than or comparable with those obtained by other researchers. However, it is noted that the number of instruments included is different and that the data sources are different despite the fact that most of these included the UIOWA sample set. The exact validation process used to assess the classification performance may be different as well. For instance, we adopted tenfold cross-validation in all our experiments, whereas Kaminskyj and

TABLE VI
CONFUSION MATRIX FOR ALL 20 INSTRUMENTS WITH TENFOLD CROSS-VALIDATION. ALL NUMBERS ARE IN PERCENTAGE

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	95	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
b=tuba	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c=trumpet	0	0	95	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d=horn	0	0	0	95	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0
e=ttrombone	0	0	0	0	90	0	0	5	0	0	0	0	0	0	0	5	0	0	0	0
f=btrombone	0	0	0	0	5	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g=violin	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
h=viola	0	0	0	0	4	8	4	72	0	0	0	0	0	0	0	4	0	4	0	4
i=bass	0	0	0	0	0	0	0	0	92	0	0	0	0	0	0	4	0	4	0	0
j=cello	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
k=sax	0	0	0	0	0	0	0	0	0	0	80	10	0	0	0	0	0	0	0	10
l=altosax	0	0	0	0	0	0	0	0	0	0	0	80	0	0	0	0	20	0	0	0
m=oboe	0	10	0	0	0	10	0	10	0	10	0	0	60	0	0	0	0	0	0	0
n=bassoon	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
o=flute	0	0	0	0	0	0	0	0	10	0	0	0	0	0	70	10	0	0	10	0
p=altoflute	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	70	20	0	0	0
q=bflute	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	80	0	0	0
r=bclarinet	0	0	0	0	0	10	0	0	10	10	0	0	0	0	10	0	0	60	0	0
s=bbclarinet	0	0	0	0	0	0	0	0	10	0	0	0	0	0	10	0	0	0	80	0
t=ebclarinet	0	0	0	0	0	0	0	0	0	0	10	10	0	0	10	20	0	0	0	50

TABLE VII
DATA SOURCES FOR THE SOLO-PHRASE EXPERIMENT

Instrument	Data sources
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

TABLE VIII
CONFUSION MATRIX FOR INSTRUMENT RECOGNITION IN SOLO PASSAGES (PERFORMANCE IN PERCENTAGE)

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

Czaszejko [19] and others used leave-one-out cross-validation instead.

Paired with a good performance level, the feature dimensionality of our approach is relatively low, with the selected feature sets having fewer than or around 20 dimensions. On the other hand, Eggink and Brown [20] 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 [21] used 62 features and selected the best 20 for real-time solo detection. Kaminskyj and Czaszejko [19] used 710 dimensions after PCA. In this paper, a 5-D set after PCA also achieved a good classification accuracy. A notable work is by Benetos *et al.* [22], where only six features were selected. However, there were only six instruments included in their study, and the scalability of the feature selection needs to be further assessed.

Although we gave such a performance list in Table IX, the comparison has to be made with a notion of care. This is particularly true for the case of instrument recognition in solo passages, as it is impossible to make fair comparison when there are no widely accepted benchmarks and researchers have used various performance CDs [8], [21].

V. 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 a number of feature-selection methods, and assessed the classification performance of the individual feature schemes, combined schemes, and selected feature subsets. A small embedding dimension of the feature space used was obtained using MDS, confirming the strong redundancy of the considered feature schemes.

For experiments on monotone music samples, a publicly available data set was used to allow for the purpose of benchmarking. Feature-ranking measures were employed, and these produced similar feature-selection outputs. Moreover, the performance of the obtained feature subsets was verified using a number of 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 and SVM-based selections. It was also demonstrated that, among the individual feature schemes, the MFCC feature scheme gave the best classification performance.

It is interesting to see that the feature schemes adopted in current research 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 quickly and accurately. The finding of an embedding dimension as low as four or five, however, may relate to the specific sound source files we used in this paper, and its scalability needs further verification.

TABLE IX
PERFORMANCE OF INSTRUMENT CLASSIFICATION COMPARED

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 [19]	19	97	93
Benetos et al. [22]	6	-	95.2
This work			
UIOWA samples	20	96.5	86.9
Solo phrases	4	-	98.4

On the other hand, in the classification of individual instruments, even the full feature set would not help much in distinguishing woodwind instruments. In fact, it was found in our experiments on solo-passage classification that some MPEG-7 features were not reliable for giving robust classification results with the current fixed segmentation of solo passages. For instance, attack time was not selected in the feature scheme, but it could become a very effective attribute with the help of onset detection. All these indicate that more research works in feature extraction and selection are 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 envelope and spectral flatness [15]. Despite some possible redundancy with the introduction of new features, it would be interesting to investigate the potential gains that can be obtained. It would also be interesting to see how the proposed approach scales with increased feature numbers and increased amount of music samples. For our future work, we intend to investigate these issues along with the use of more live recorded music data and also 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. 22nd ACM SIGIR Int. Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 176–182.
[2] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proc. IEEE*, vol. 92, no. 4, pp. 712–729, Apr. 2004.
[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.
[4] T. Lidy and A. Rauber, "Evaluation of feature extractors and psychoacoustic transformations for music genre classification," in *Proc. 6th Int. Conf. ISMIR*, Sep. 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 Comput. Corp., Tech. Rep. CRL 99/4, 1999.
[6] A. Eronen, "Comparison of features for musical instrument recognition," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 19–22.
[7] G. Agostini, M. Longari, and E. Poolastri, "Musical instrument timbres classification with spectral features," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 1, pp. 5–14, 2003.
[8] S. Essid, G. Richard, and B. David, "Efficient musical instrument recognition on solo performance music using basic features," presented at the Audio Engineering Society 25th Int. Conf., London, U.K., 2004, Paper 2–5. 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 Syst.*, vol. 7, no. 1, pp. 2–10, Jan. 1999.
[10] H. G. Kim, N. Moreau, and T. Sikora, "Audio classification based on MPEG-7 spectral basis representations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 716–725, May 2004.
[11] J. Aucouturier and F. Pachet, "Scaling up music playlist generation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2002, vol. 1, 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. IEEE Int. Conf. Multimedia Expo*, 2003, vol. 3, pp. 397–400.
[13] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–22, Jul. 2006.
[14] A. Divakaran, R. Regunathan, Z. Xiong, and M. Casey, "Procedure for audio-assisted browsing of news video using generalized sound recognition," *Proc. SPIE*, vol. 5021, pp. 160–166, 2003.
[15] ISO/IEC Working Group, *MPEG-7 Overview*, 2004. accessed 8.2.2007. [Online]. Available: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
[16] A. T. Lindsay and J. Herre, "MPEG-7 audio—An overview," *J. Audio Eng. Soc.*, vol. 49, no. 7/8, pp. 589–594, Jul./Aug. 2001.
[17] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7," in *Proc. Int. Comput. Music Conf.*, 2000, pp. 166–169.
[18] J. C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1064–1072, Mar. 2001.
[19] I. Kaminskyj and T. Czaszejko, "Automatic recognition of isolated monophonic musical instrument sounds using kNNC," *J. Intell. Inf. Syst.*, vol. 24, no. 2/3, pp. 199–221, Mar. 2005.
[20] J. Eggink and G. J. Brown, "Instrument recognition in accompanied sonatas and concertos," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. IV, pp. 217–220.
[21] A. A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Proc. 7th Int. Conf. Digital Audio Effects*, 2004, 222–226. [Online]. Available: http://dafx04.na.infn.it/
[22] E. Benetos, M. Kotti, and C. Kotropoulos, "Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection," in *Proc. ICASSP*, 2006, vol. V, pp. 221–224.
[23] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
[24] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.
[25] M. Grimaldi, P. Cunningham, and A. Kokaram, "An evaluation of alternative feature selection strategies and ensemble techniques for classifying music," School Comput. Sci. and Inf., Trinity College Dublin, Dublin, Ireland, Tech. Rep. TCD-CS-2003-21, 2003.
[26] J. Qinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
[27] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[28] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1/2, pp. 273–324, Dec. 1997.
[29] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[30] C. Atkeson, A. Moore, and S. Schaal, "Locally weighted learning," *Artif. Intell. Rev.*, vol. 11, no. 1–5, pp. 11–73, Feb. 1997.
[31] IPEM, *IPEM-Toolbox*. accessed 10/9/2005. [Online]. Available: http://www.ipem.ugent.be/Toolbox

[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. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 737–747, Jun. 2001.

[34] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

[35] K. D. Martin and Y. E. Kim, "Musical instrument identification: A pattern-recognition approach," *J. Acoust. Soc. Amer.*, vol. 103, no. 3, p. 1768, 1998.

Jeremiah D. Deng (M'01) received the B.Eng. degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1989, and the M.Eng. and Ph.D. degrees from the South China University of Technology, Guangzhou, ,China in 1992 and 1995, respectively.

He was a Lecturer with the South China University of Technology during 1995–1999 before he joined the University of Otago, Dunedin, New Zealand, as a Postdoctoral Fellow. He is currently a Senior Lecturer with the Department of Information Science, University of Otago. He has published more than 40 refereed papers. His research interests include digital signal processing, multimedia information retrieval, neural networks, and telecommunications.

Christian Simmermacher received the Diploma in information and knowledge management from the University of Applied Sciences, Darmstadt, Germany, in 2003, and the M.Sc. degree in information science from Otago University, Dunedin, New Zealand, in 2006.

He is currently with IBM Germany, Herrenberg, Germany. He is also currently working as a System Consultant for the German Army in the Herkules project. His research works covered the topics of metadata management for media assets and instrument detection in music. **AQ6**

Stephen Cranefield received the B.Sc.(Hons.) degree in mathematics from the University of Otago, Dunedin, New Zealand, in 1985, and the Ph.D. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K., in 1991. **AQ7**

He was a Lecturer with Massey University, Wellington, New Zealand, from 1993 to 1994 before moving to the University of Otago where he is currently an Associate Professor. He has published more than 90 refereed publications. His research interest is mainly focused on multiagent systems and electronic institutions. **AQ8**

Dr. Cranefield is in the Editorial Board of the *Knowledge Engineering Review* journal.