EURASIP Journal on Applied Signal Processing 2003:1, 5–14

Musical Instrument Timbres Classification with Spectral Features

Giulio Agostini
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, 20135 Milano, Italy
Email: guilio@despammed.com

Maurizio Longari
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, 20135 Milano, Italy
Email: longari@dsi.unimi.it

Emanuele Pollastri
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, 20135 Milano, Italy
Email: pollastri@dsi.unimi.it

Received 10 May 2002 and in revised form 29 August 2002

A set of features is evaluated for recognition of musical instruments out of monophonic musical signals. Aiming to achieve a compact representation, the adopted features regard only spectral characteristics of sound and are limited in number. On top of these descriptors, various classification methods are implemented and tested. Over a dataset of 1007 tones from 27 musical instruments, support vector machines and quadratic discriminant analysis show comparable results with success rates close to 70% of successful classifications. Canonical discriminant analysis never had momentous results, while nearest neighbours performed on average among the employed classifiers. Strings have been the most misclassified instrument family, while very satisfactory results have been obtained with brass and woodwinds. The most relevant features are demonstrated to be the inharmonicity, the spectral centroid, and the energy contained in the first partial.

Keywords and phrases: timbre classification, content-based audio indexing/searching, pattern recognition, audio features extraction.

1. INTRODUCTION

This paper addresses the problem of musical instrument classification from audio sources. The need for this application strongly arises in the context of multimedia content description. A great number of commercial applications will be available soon, especially in the field of multimedia databases, such as automatic indexing tools, intelligent browsers, and search engines with querying by content capabilities.

The goal of automatic music-content understanding and description is not new and it is traditionally divided into two subtasks: pitch detection, or the extraction of score-like attributes from an audio signal (i.e., notes and durations), and sound-source recognition, or the description of sounds involved in an excerpt of music [1]. The former has received a lot of attention and some recent experiments are described in [2, 3]; the latter has not been studied so much because of the lack of knowledge about human perception and cognition of sounds. This work belongs to the second area and it is devoted to a more modest goal, but important nevertheless, automatic timbre classification of audio sources containing no more than one instrument at a time (source must be monotimbral and monophonic).

Focusing on this area, the forthcoming MPEG-7 standard should provide a list of metadata for multimedia content [4], nevertheless, two important aspects still need to be explored further. First, the best features for a particular task must be identified. Then, once obtained a set of descriptors, some classification algorithms should be employed to organize metadata in meaningful categories. All these facets will be considered by the present work with the objective of automatic timbres classification for sound databases.

This paper is organized as follows. First, we give some background information on the notion of timbre and previous related works; then, some details about feature properties and calculation are presented. A brief description of various classification techniques is followed by the experiments. Finally, results are presented and compared to previous studies on the same topic. Discussion and further work close the paper.

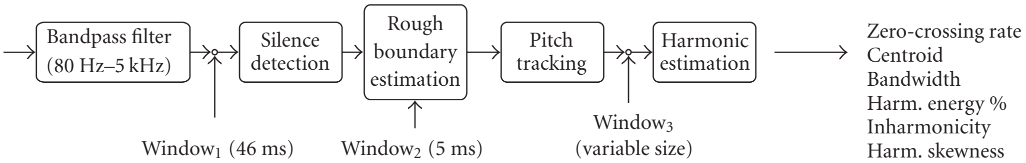

Figure 1: Description of the feature extraction process.

2. BACKGROUND

Timbre differs from the other sound attributes; namely, pitch, loudness, and duration, because it is ill-defined; in fact, it cannot be directly associated with a particular physical quantity. The American National Standards Institute (ANSI) defines timbre as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" [5]. The uncertainty about the notion of timbre is reflected by the huge amount of studies that have tackled this problem. Since the first studies by Grey [6], it was clear that we are dealing with a multidimensional attribute, which includes spectral and temporal features. Therefore, early works on timbre recognition focused on the exploration of possible relationships between the perceptual and the acoustic domains. The first experiments on sound classification are illustrated in [7, 8, 9] where a limited number of musical instruments (eight instruments or less) has been recognized, implementing a basic set of features. Other works explored issues about the relationship between acoustic features and sound properties [10, 11], justifying their choice in terms of musical relevance, brightness, spectral synchronicities, harmonicity, and so forth. Recently, the diffusion of multimedia databases has brought to the forth problem of musical instrument identification out of a fragment of audio signal. In this context, deep investigations on sound classification as a pattern recognition problem began to appear in the last few years [12, 13, 14, 15, 16, 17]. These works emphasized the importance of testing different classifiers and set of features with datasets of dimension comparable to real world applications. Further works related to timbre classification have dealt with the more general problem of audio segmentation [18, 19], especially with the purpose of automatic (video) scene segmentation [20]. Finally, the introduction of content management applications like the ones envisioned by MPEG-7 boosted the interest in the topic [4, 21].

3. FEATURE EXTRACTION

A considerable number of features is currently available in the literature, each one describing some aspects of audio content [22, 23]. In the digital domain, features are usually calculated from a window of samples, which is normally very short compared to the total duration of a tone. Thus, we must face the problem of summarizing their temporal evolution into a small set of values. Mean, standard deviation, skewness, and autocorrelation have been the preferred strategies for their simplicity, but more advanced methods like hidden Markov models could be employed, as illustrated in [21, 22]. By combining these time-spanning statistics with the known features, an impressive number of variables can be extracted from each sound. The researcher, though, has to carefully select them in order to both keep the time required for the extraction to a minimum and, more importantly, to prevent from incurring into the so-called curse of dimensionality. This fanciful term refers to a well-known result of classification theory [24] which states that, as the number of variables grows, in order to maintain the same error rate, the classifier has to be trained with an exponentially growing training set. The process of feature extraction is crucial; it should perform efficient data reduction while preserving the appropriate amount of information. Thus, sound analysis techniques must be tailored to the temporal and spectral evolution of musical signals. As it will be demonstrated in Section 6, a set of features related mainly to the harmonic properties of sounds allows a simplified representation of data. However, lacking features for the discrimination between sustained sounds and percussive sounds, a classification solely based on spectral properties has some drawbacks (see Section 7 for details).

The extraction of descriptors relies on a number of preliminary steps: temporal segmentation of the signal, detection of the fundamental frequency, and the estimation of the harmonic structure (Figure 1).

3.1. *Audio segmentation*

The aim of the first stage is twofold. First of all, the audio signal must be segmented into a sequence of meaningful events. We do not make any assumptions about the content of each event, which corresponds to an isolated tone in the ideal case. Subsequently, a decision based on the pitch estimation is taken for a fine adjustment of event boundaries. The output of this stage is a list of nonsilent events (starting and ending points) and estimated pitch values.

In the experiment reported in this paper, we assume to deal with audio signals characterized by a low level of noise and a good dynamic range. Therefore, a simple procedure based on energy evaluation is expected to satisfactorily perform in the segmentation task. The signal is first processed with a bandpass Chebyshev filter of order five; cutoff frequencies are set to 80 Hz to filter out noise due to unwanted vibrations (for instance, oscillation of the microphone stand) and 5000 Hz, corresponding to E8 in a tempered musical scale. After windowing the signal (46 ms Hamming), an root mean square (RMS)-energy curve is computed with the same frame size. By comparing the energy to an absolute threshold empirically set to −50 dB (0 dB being the full scale reference value), we find out a rough estimate of the boundaries of the events. A finer analysis is then conducted with a 5-ms frame to determine actual on/offsets; in particular, we look for a 6-dB step around every rough estimate. Through pitch detection, we achieve a refinement of signal segmentation, identifying notes that are not well defined by the energy curve or that are possibly played legato. Pitch is also input to the calculation of some spectral features. The pitch-tracking algorithm employed follows the one presented in [25], so it will not be described here. The output of the pitch tracker is the average value (in hertz) of each note hypothesis, a frame-by-frame value of pitch and a confidence value that measures the uncertainty of the estimate.

3.2. Spectral features

We collect a total of 18 descriptors for each tone isolated through the procedure just described. More precisely, we compute mean and standard deviation of 9 features over the length of each tone. The zero-crossing rate is measured directly from the waveform as the number of sign inversions within a 46 ms window. Then, the harmonic structure of the signal is evaluated through a short-time Fourier analysis with half-overlapping windows. The size of the analysis window is variable in order to have a frequency resolution of at least 1/24 of octave, even for the lowest tones (1024–8192 samples, for tones sampled at 44100 Hz). The signal is first analyzed at a low-frequency resolution; the analysis is repeated with finer resolutions until a sufficient number of harmonics is estimated. This process is controlled by the pitch-tracking algorithm [25]. From the harmonic analysis, we calculate spectral centroid and bandwidth according to the following equations:

$$\begin{aligned}\text{Centroid} &= \frac{\sum_{f=f_{\min}}^{f_{\max}} f \cdot E(f)}{\sum_{f=f_{\min}}^{f_{\max}} E(f)},\\ \text{Bandwidth} &= \frac{\sum_{f=f_{\min}}^{f_{\max}} |\text{centroid} - f| \cdot E(f)}{\sum_{f=f_{\min}}^{f_{\max}} E(f)},\end{aligned} \tag{1}$$

where $f_{\min} = 80\,\text{Hz}$ and $f_{\max} = 5000\,\text{Hz}$, and $E(f)$ is the energy of the spectral component at frequency f.

Since several sounds slightly deviate from the harmonic rule, a feature called inharmonicity is measured as a cumulative distance between the first four estimated partials (p_i) and their theoretical values ($i \cdot f_0$, where f_0 is the fundamental frequency of the sound),

$$\text{Inharmonicity} = \sum_{i=1}^{4} \frac{|p_i - i \cdot f_0|}{i \cdot f_0}. \tag{2}$$

The percentage of energy contained in each one of the first four partials is calculated for bins 1/12 oct wide, providing four different features.

Finally, we introduce a feature obtained by combining the energy confined in each partial and its respective inharmoncity

$$\text{Harmonic energy skewness} = \sum_{i=1}^{4} \frac{|p_i - i \cdot f_0|}{i \cdot f_0} \cdot E_{p_i}, \tag{3}$$

where E_{p_i} is the percentage of energy contained in the respective partial.

4. CLASSIFICATION TECHNIQUES

In this section, we provide a brief survey on the most popular classification techniques, comparing different approaches. As an abstract task, pattern recognition aims at associating a vector $\mathbf{y}$ in a p-dimensional space (the feature space) to a class, given a dataset (or training set) of N vectors $\mathbf{d}_i$. Since each of these observations belong to a known class, among the c available, this is said to be a supervised classification. In our instance of the problem, the features extracted are the dimensions, or variables, and the instrument labels are the classes. The vector $\mathbf{y}$ represents the tone played by an unknown musical instrument.

4.1. Discriminant analysis

The multivariate statistical approach to the question [26] has a long tradition of research. Considering $\mathbf{y}$ and $\mathbf{d}_i$ as realizations of random vectors, the probability of a misclassification of a classifier g can be expressed as a function of the probability density functions (PDFs) $f_i(\cdot)$ of each class

$$\gamma_g = 1 - \sum_{i=1}^{c} \left(\pi_i \int_{\mathbb{R}^p} f_i(\mathbf{y})d\mathbf{y}\right), \tag{4}$$

where π_i is the a priori probability that an observation belongs to the ith class. It can also be proven that the optimal classifier, which is the classifier that minimizes the error rate, is the one that associates to the ith class every vector $\mathbf{y}$ for which

$$\pi_i f_i(\mathbf{y}) > \pi_j f_j(\mathbf{y}), \quad \forall i \neq j. \tag{5}$$

Unfortunately, PDFs $f_i(\cdot)$ are generally unknown. Nonetheless, we can make assumptions about the distributions of the classes and estimate the necessary parameters to obtain a good guess of those functions.

4.1.1 Quadratic discriminant analysis (QDA)

This technique starts from the working hypothesis that classes have multivariate normal PDFs. The only parameters characterizing those distributions are the mean vectors $\boldsymbol{\mu}_i$ and the covariance matrices $\boldsymbol{\Sigma}_i$. We can easily estimate them by computing the traditional sample statistics

$$\begin{aligned}\mathbf{m}_i &= \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{d}_{ij},\\ \mathbf{S}_i &= \frac{1}{N_i - 1}\sum_{j=1}^{N_i} (\mathbf{d}_{ij} - \mathbf{m}_i)(\mathbf{d}_{ij} - \mathbf{m}_i)',\end{aligned} \tag{6}$$

using the N_i observations $\mathbf{d}_{ij}$ available for the ith class from the training sequence. It can be shown that, in this case, the hypersurfaces delimiting the regions of classification—in which the associated class is the same—are quadratic forms, hence the name of the classifier.

Although this is the optimal classifier for normal mixtures, it could lead to suboptimal error rates in practical cases for two reasons. First, classes may depart sensibly from the assumption of normality. A more subtle source of errors is the fact that, with this method, the actual distributions remain unknown, since we only have the best estimates of them, based on a finite training set.

4.1.2 *Canonical discriminant analysis*

The canonical discriminant analysis (CDA) is a generalization of the linear discriminant analysis which separates two classes ($c = 2$) in a plane ($p = 2$) by means of a line. This line is found by maximizing the separation of the two one-dimensional distributions that result from the projection of the two bivariate distributions on the direction normal to the line of separation sought.

In a p-dimensional space, using a similar criterion, we can separate $c \geq 2$ classes with hyperplanes by maximizing, with respect to a generic vector $\mathbf{a}$, the figure of merit

$$D(\mathbf{a}) = \frac{\mathbf{a}'\mathbf{S}_B\mathbf{a}}{\mathbf{a}'\mathbf{S}_W\mathbf{a}}, \tag{7}$$

where

$$\mathbf{S}_B = \frac{1}{N}\sum_{j=1}^{c} N_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})' \tag{8}$$

is the between-class scatter matrix, and

$$\mathbf{S}_W = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i}(\mathbf{d}_{ij} - \mathbf{m}_i)(\mathbf{d}_{ij} - \mathbf{m}_i)' \tag{9}$$

is the within-class scatter matrix, $\mathbf{m}$ being the sample mean of all the observations, and N the total number of observations. Equivalent to QDA from the point of view of computational complexity, CDA has proven to perform better when there are few samples available, because it is less sensitive to overfitting. CDA and QDA are identical (i.e., optimal) rules under homoscedasticity conditions. Thus, if the underlying covariance matrices are quite different, QDA has lower error rates. QDA is also preferred in presence of long tails and pronounced kurtosis, whereas a moderate skewness suggests to use CDA.

4.2. *k-nearest neighbours (k-NN)*

This is one of the most popular nonparametric techniques in pattern recognition. It does not require any knowledge about the distribution of the samples and it is quite easy to implement. In fact, this method classifies $\mathbf{y}$ as belonging to the class which is most frequent among its k-nearest observations. Thus, only two parameters are needed: a distance metric and the number of nearest samples considered (k). An important drawback is its poor ability to abstract from data since only local information is taken into account.

4.3. *Support vector machines*

The support vector machines (SVM) are a recently developed approach to the learning problem [27]. The aim is to find the hyperplane that best separates observations belonging to different classes. This is done by satisfying a generalization bound which maximizes the geometric margin between the sample data and the hyperplane, as briefly detailed below.

Suppose we have a set of linearly separable training samples $\mathbf{d}_1, \dots, \mathbf{d}_N$, with $\mathbf{d}_i \in \mathbb{R}^p$. We refer to the simplified binary classification problem (two classes, $c = 2$), in which a label $l_i \in \{-1, 1\}$ is assigned to the ith sample, indicating the class they belong to. The hyperplane $f(\mathbf{y}) = (\mathbf{w} \cdot \mathbf{y}) + b$ that separates the data can be found by minimizing the 2-norm of the weight vector $\mathbf{w}$,

$$\min_{\mathbf{w},b}\langle \mathbf{w} \cdot \mathbf{w}\rangle \tag{10}$$

subject to the following class separation constraints:

$$l_i(\langle \mathbf{w} \cdot \mathbf{d}_i\rangle + b) \geq 1, \quad 1 \leq i \leq N. \tag{11}$$

This approach is called maximal margin classifier. The optimal solution can be viewed in a dual form by applying the Lagrange theory and imposing the conditions of stationariness. The objective and decision functions can thus be written in terms of the Lagrange multipliers α_i as

$$\begin{gathered} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} l_i l_j \alpha_i \alpha_j \langle \mathbf{d}_i \cdot \mathbf{d}_j\rangle, \\ f(\mathbf{y}) = \sum_{i=1}^{N} l_i \alpha_i \langle \mathbf{d}_i \cdot \mathbf{y}\rangle + b. \end{gathered} \tag{12}$$

The support vectors are defined as the input samples $\mathbf{d}_i$ for which the respective Lagrange multiplier α_i is nonzero, so they contain all the information needed to reconstruct the hyperplane. Geometrically, they are the closest samples to the hyperplane to lie on the border of the geometric margin.

In case the classes are not linearly separable, the samples are projected through a nonlinear function $\Phi(\cdot)$ from the input space Y in a higher-dimensional space (with possibly infinite dimensions), which we will call the transformed space[1] T. The transformation $\Phi(\mathbf{y}) : Y \to T$ has to be a nonlinear function so that the transformed samples can be linearly separable. Since the high number of dimensions increases the computational effort, it is possible to introduce the *kernel functions* $K(\mathbf{y}, \mathbf{z}) = \langle \Phi(\mathbf{y}) \cdot \Phi(\mathbf{z})\rangle$, which implicitly define the transformation $\Phi(\cdot)$ and allow to find the solution in the transformed space T by making simpler calculations in the input space Y. The theory does not grant that the

[1] For the sake of clarity, we will avoid the traditional name "feature space."

Table 1: Taxonomy of the instruments employed in the experiments.

Pizzicati			Sustained		
Piano et al.	Rock strings	Pizz. strings	Strings	Woodwinds	Brass
Piano	Electric bass	Violin pizzicato	Violin bowed	Flute	C trumpet
Harpsichord	Elect. bass slap	Viola pizzicato	Viola bowed	Organ	French horn
Classic guitar	Electric guitar	Cello pizzicato	Cello bowed	Accordion	Tuba
Harp	Dist. elect. guitar	Doublebass pizz.	Doublebass bowed	Bassoon	
				Oboe	
				English horn	
				E♭ clarinet	
				Sax	

best linear hyperplane can always be found, but, in practice, a solution can be heuristically obtained. Thus, the problem is now to find a kernel function that well separates the observations. Not just any function is a kernel function; it must be symmetric, it must satisfy the Cauchy-Schwartz inequality, and must satisfy the condition imposed in Mercer's theorem. The simplest example of a kernel function is the dot kernel, which maps the input space directly into the transformed space. Radial basis functions (RBF) and polynomial kernels are widely used in image recognition, speech recognition, handwritten digit recognition, and protein homology detection problems.

5. EXPERIMENT

The adopted dataset has been extracted by the MUMS (McGill University Master Samples) CDs [28], which is a library of isolated sample tones from a wide number of musical instruments, played with several articulation styles and covering the entire pitch range. We considered 30 musical instruments ranging from orchestral sounds (strings, woodwinds, brass) to pop/electronic instruments (bass, electric, and distorted guitar). An extended collection of musical instrument tones is essential for training and testing classifiers for two distinct reasons. First, methods that require an estimate of the covariance matrices, namely, QDA and CDA, must compute it with at least $p + 1$ linearly independent observations for each class, p being the number of features extracted, so that they are definite positive. In addition, we need to avoid the curse of dimensionality discussed in page 6, therefore a rich collection of samples brings the expected error rate down. It follows from the first observation that we could not include musical instruments with less than 19 tones in the training set. This is why we collapsed the family of saxophones (alto, soprano, tenor, baritone) to a single instrument class.[2] Having said that, the total number of musical instruments considered was 27, but the classification results reported in Section 6 can be claimed to hold for a set of 30 instruments (Table 1).

The audio files have been analyzed by the feature extraction algorithms. If the accuracy of a pitch estimate is below a predefined threshold, the corresponding tone is rejected from the training set. Following this procedure, the number of tones accepted for training/testing is 1007 in total. Various classification techniques have been implemented and tested: CDA, QDA, k-NN, and SVM. k-NN has been tested with $k = 1, 3, 5, 7$ and with 3 different distance metrics (1-norm, 2-norm, 3-norm). In one experiment, we modified the input space through a kernel function. For SVM, we adopted a software tool developed at the Royal Holloway University of London [29]. A number of kernel functions has been considered (dot product, simple polynomial, RBF, linear splines, regularized Fourier). Input values have been normalized independently and we chose a multiclass classification method that trains $c(c - 1)/2$ binary classifiers, where c is the number of instruments. Therefore, recognition rates in the classification of instrument families have been calculated by grouping results from the recognition of individual instruments. All error rates estimates reported in Section 6 have been computed using a leave-one-out procedure.

6. RESULTS

The experiments illustrated have been evaluated by means of overall success rate and confusion matrices. In the first case, results have been calculated as the ratio of estimated and actual stimuli. Confusion matrices represent a valid method for inspecting performances from a qualitative point of view. Although we put the emphasis on the instrument level, we have also grouped instruments belonging to the same family (strings, brass, woodwinds and the like), extending Sachs taxonomy [30] with the inclusion of rock strings (deep bass, electric guitar, distorted guitar). Figure 2 provides a graphical representation of the best results both at the instrument level (17, 20, and 27 instruments) and at the family level (pizzicato-sustained, instrument family).

SVM with RBF kernel was the best classifier in the recognition of individual instruments, with a success rate

[2] We observe that the recognition of the single instrument within the sax class can be easily accomplished by inspecting the pitch, since the ranges do not overlap.

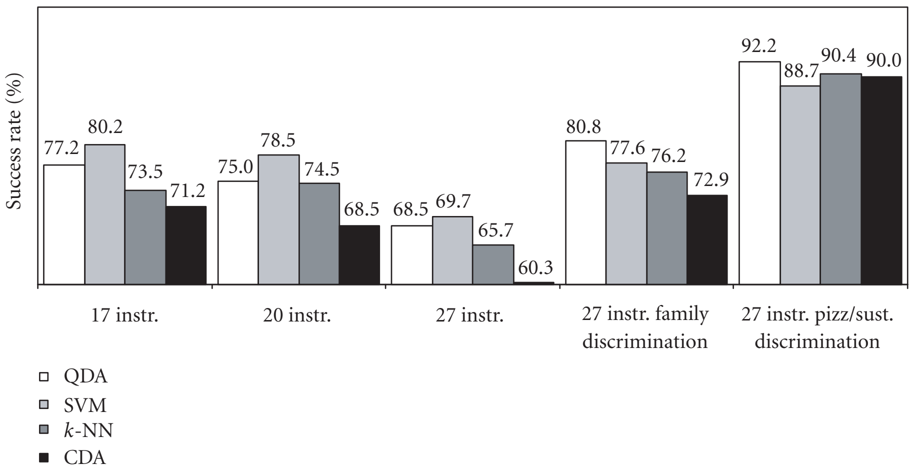

Figure 2: Graphical representation of the success rates for each experiment.

of 69.7%, 78.6%, and 80.2% for, respectively, 27, 20, and 17 instruments. In comparison with the work by Marques and Moreno [15], where 8 instruments were recognized with an error rate of 30%, the SVM implemented in our experiments had an error rate of 19.8% in the classification of 17 instruments. The second best score was achieved by QDA, with success rates close to SVM's performances. In the case of instrument family recognition and sustain/pizzicato classification, QDA overcame all other classifiers with a success rate of 81%. Success rates with SVM at the family and pizzicato/sustained levels should be carefully evaluated since we did not train a new SVM for each family (i.e., grouping instruments by family or pizzicato/sustained). Thus, we have to consider results for pizzicato/sustained discrimination for this classifier as merely indicative although success rates with all classifiers are comparable for this task.

CDA never obtained momentous results, ranging from 71.2% with 17 instruments to 60.3% with 27 instruments. In spite of their simplicity, k-NN performed quite close to QDA. Among the k-NN classifiers, 1-NN with 1-norm distance metric obtained the best performance. Since the k-NN was employed in a number of experiments, we observe that our results are similar to those previously reported, for example, in [31]. Using a kernel function to modify the input space did not bring any advantage (71% with kernel and 74.5% without kernel for 20 instruments).

A deeper analysis of the results achieved with SVM and QDA (see Figures 3, 4, 5, 6) showed that strings have been the most misclassified family with 39.52% and 46.75% of individual instruments identified correctly on average, respectively, for SVM and QDA. Leaving out strings samples, the success rates for the remaining 19 instruments grow up to some 80% for the classification of individual instruments. Since this behaviour has been registered for both pizzicati and sustained strings, we should conclude that our features are not suitable for describing such instruments. In particular, SVM classifiers seem to be unable to recognize the doublebass and the pizzicato strings, for which, results have been as low as some 7% and 30%; instead, sustained strings have been identified correctly in 64% of cases, conforming to the overall rate. QDA classifiers did not show a considerable difference in performance between pizzicato and sustained strings. Moreover, most of the misclassifications have been within the same family. This fact explains the slight advantage of QDA in the classifications at the family level.

The recognition of woodwinds, brass, and rock strings has been very successful (94%, 96%, 89% with QDA), without noticeable differences between QDA and SVM. Misclassifications within these families reveal strong and well-known subjective evidence. For example, basoon has been estimated as tuba (21% with QDA), oboe as flute (11% with QDA), and deep bass as deep bass slap (24% with QDA). The detection of stimuli from the family of piano and other instruments is definitely more spread around the correct family, with success rates for the detection of this family close to 70% with SVM and to 64% with QDA.

We have also calculated a list of the most relevant features through the forward selection procedure detailed in [32]. The values reported are the normalized versions of the statistics on which the procedure is based, and can be interpreted as the amount of the information added by each feature. They cannot be strictly decreasing because a feature might bring more information only jointly with other features. For 27 instruments, the most informative feature has been the mean of the inharmonicity, followed by the mean and standard deviation of the spectral centroid and the mean of the energy contained in the first partial (see Table 2).

In one of our experiments, we have also introduced a machine-built decisional tree. We used a hierarchical clustering algorithm [33] to build the structure. CDA or QDA methods have been employed at each node of the hierarchy. Even with these techniques, though, we could not improve the error rates, thus confirming the previous findings [13].

Stimulus input / Recognize as	Hamburg steinway	Harpsichord	Classic guitar	Harp	Deep electric bass	Deep electric bass slap	Electric guitar	Distorted electric guitar	Violin pizz.	Viola pizz.	Cello pizz.	Doublebass pizz.
Hamburg steinway	64		8							15		
Harpsichord		76					2					
Classic guitar	2		42						6	10		14
Harp	5			58	6	16				1		
Deep electric bass	3			17	71	21						
Deep electric bass slap	2			25	24	63						
Electric guitar		12					94					
Distorted electric guitar			2					82	6			
Violin pizz.	3								39	21		
Viola pizz.	3		9						22	18	7	
Cello pizz.			3					6	11	12	64	21
Doublebass pizz.	2		6							7	29	64
Family success (%)	64.00				88.64				79.04			
Pizzicato success (%)	91.27											

Figure 3: Confusion matrix for the classification of individual instruments in the family of pizzicati with QDA.

Stimulus input / Recognize as	Violin bowed	Viola bowed	Cello bowed	Doublebass bowed	Flute	B. Plenum organ	Accordion	Bassoon	Oboe	English horn	E♭ clarinet	Sax	C trumpet	French horn	Tuba
Violin bowed	38	39													
Viola bowed	30	37				4									
Cello bowed			35	11											
Doublebass bowed			18	44								2			
Flute					94				11			5			
B. Plenum organ	2					91			4						
Accordion			8				100								
Bassoon			1					68				5			
Oboe					6	5			79			12			
English horn	6									80	4				
E♭ clarinet	8	4								12	91		4		
Sax	3	6						6	7			73	4		
C trumpet	6	7								8	4		92		
French horn								6						100	
Tuba								21							95
Family success (%)	62.92				94.08								95.82		
Sustained success (%)	93.07														

Figure 4: Confusion matrix for the classification of individual instruments in the family of sustained with QDA.

7. DISCUSSION AND FURTHER WORK

A thorough evaluation of the resulting performances illustrated in Section 6 reveals the power of SVM in the task of timbre classification, thus confirming the successful results in other fields (e.g., face detection, text classification). Furthermore, in our experiments, we employed widely used kernel functions, so there is a room for improvement adopting dedicated kernels. However, QDA performed similarly in the recognition of individual instruments with errors closer to the way human classify sounds. It was highlighted that much of the QDA errors are within the correct family, while

Recognize as \ Stimulus input	Hamburg steinway	Harpsichord	Classic guitar	Harp	Deep electric bass	Deep electric bass slap	Electric guitar	Distorted electric guitar	Violin pizz.	Viola pizz.	Cello pizz.	Doublebass pizz.
Hamburg steinway	69		3	4					4	14	3	4
Harpsichord		75										
Classic guitar			64		14	13						
Harp	2			61				8	7	10	7	11
Deep electric bass			6		50	23				3		
Deep electric bass slap			19		32	63						
Electric guitar		6					96					
Distorted electric guitar								76	4		3	4
Violin pizz.	3		3						33	17	3	
Viola pizz.	3		3	7			2		41	34	10	4
Cello pizz.				7				0	4	10	41	22
Doublebass pizz.	11			2				8				7
Family success (%)	69.41				85.24				57.09			
Pizzicato success (%)	86.29											

FIGURE 5: Confusion matrix for the classification of individual instruments in the family of pizzicati with SVM.

Recognize as \ Stimulus input	Violin bowed	Viola bowed	Cello bowed	Doublebass bowed	Flute	B. Plenum organ	Accordion	Bassoon	Oboe	English horn	Eb clarinet	Sax	C trumpet	French horn	Tuba
Violin bowed	64	29									3		3	3	
Viola bowed	27	60				2									
Cello bowed			70	18			5								
Doublebass bowed				7											
Flute					100							2			
B. Plenum organ		2				96			6						
Accordion				5			85								
Bassoon								81		3		5			3
Oboe						2			84			10			
English horn							3	9		92					
E♭ clarinet	2	2									81		3	7	
Sax	2								9			80			
C trumpet											6		94	3	
French horn		2									9			83	
Tuba								6							94
Family success (%)	68.60				93.70								91.51		
Sustained success (%)	91.04														

FIGURE 6: Confusion matrix for the classification of individual instruments in the family of sustained with SVM.

SVM show errors scattered throughout the confusion matrices. Since QDA is the optimal classifier under multivariate normality hypotheses, we should conclude that the features we extracted from isolated tones follow such distribution. To validate this hypothesis, a series of statistical tests are undergoing on the dataset.

As it was anticipated, sounds that exhibit a predominant percussive nature are not well characterized by a set of features solely based on spectral properties, while sustained sounds like brass are perfectly tailored. Our experiments have demonstrated that classifiers are not able to overcome this difficulty. Moreover, the closeness of performances between

Table 2: Most discriminating features for 27 instruments.

Feature Name	Score
Inharmonicity mean	1.0
Centroid mean	0.202121
Centroid standard deviation	0.184183
Harmonic energy percentage (partial 0) mean	0.144407
Zero-crossing mean	0.130214
Bandwidth standard deviation	0.141585
Bandwidth mean	0.1388
Harmonic energy skewness standard deviation	0.130805
Harmonic energy percentage (partial 2) standard deviation	0.116544

k-NN and SVM indicates that the choice of features is more critical than the choice of a classification method. However, that may be—beside a set of spectral features, it is important to introduce temporal descriptors of sounds—like the log attack slope or similar.

The method employed in our experiments to extract features out of a tone (i.e., mean and standard deviation) does not consider the time-varying nature of sounds known as articulation. If the multivariate normality hypotheses were confirmed, a suitable model of articulation is the continuous hidden Markov model, in which the PDFs of each state is Gaussian [21].

The experiments described so far has been conducted on real acoustic instruments with relatively little influence of the reverberant field. A preliminary test with performances of trumpet and trombone has shown that our features are quite robust against the effects of room acoustics. The only weakness is their dependence from the pitch, which can be reliably estimated out of monophonic sources only. We are planning to introduce novel harmonic features that are independent of pitch estimation.

As a final remark, it is interesting to compare our results with human performances. In a recent paper [34], 88 conservatory students were asked to recognize 27 musical instruments out of a number of isolated tones randomly played by a CD player. An average of 55.7% of tones has been correctly classified. Thus, timbre recognition by computer model is able to exceed human performance under the same conditions (isolated tones).

ACKNOWLEDGMENTS

Authors are grateful to Prof. N. Cesa Bianchi, Ryan Rifkin, and Alessandro Conconi for the fruitful discussions about SVM and pattern classification. Portions of this work were presented at the Multimedia Signal Processing 2001 IEEE Workshop and the Content-Based Multimedia Indexing 2001 IEEE Workshop.

REFERENCES

[1] G. Peeters, S. McAdams, and P. Herrera, “Instrument sound description in the context of MPEG-7,” in *Proc. International Computer Music Conference*, pp. 166–169, Berlin, Germany, August-September 2000.

[2] T. Virtanen and A. Klapuri, “Separation of harmonic sounds using linear models for the overtones series,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Orlando, Fla, USA, May 2002.

[3] P. J. Walmsley, “Polyphonic pitch tracking using joint bayesian estimation of multiple frame parameters,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, October 1999.

[4] Moving Pictures Experts Group, “Overview of the MPEG-7 standard,” Document ISO/IEC JTC1/SC29/WG11 N4509, Pattaya, Thailand, December 2001.

[5] American National Standards Institute, *American National Psychoacoustical Terminology. S3.20*, Acoustical Society of America (ASA), New York, NY, USA, 1973.

[6] J. M. Grey, “Multidimensional perceptual scaling of musical timbres,” *Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.

[7] P. Cosi, G. De Poli, and P. Prandoni, “Timbre characterization with Mel-Cepstrum and neural nets,” in *Proc. International Computer Music Conference*, pp. 42–45, Aarhus, Denmark, 1994.

[8] B. Feiten and S. Günzel, “Automatic indexing of a sound database using self-organizing neural nets,” *Computer Music Journal*, vol. 18, no. 3, pp. 53–65, 1994.

[9] I. Kaminskyj and A. Materka, “Automatic source identification of monophonic musical instrument sounds,” in *Proc. IEEE Int. Conf. Neural Networks*, vol. 1, pp. 189–194, Perth, Australia, November 1995.

[10] S. Dubnov, N. Tishby, and D. Cohen, “Polyspectra as measures of sound texture and timbre,” *Journal of New Music Research*, vol. 26, no. 4, pp. 277–314, 1997.

[11] S. Rossignol, X. Rodet, J. Soumagne, J. L. Colette, and P. Depalle, “Automatic characterisation of musical signals: Feature extraction and temporal segmentation,” *Journal of New Music Research*, vol. 28, no. 4, pp. 281–295, 1999.

[12] J. C. Brown, “Musical instrument identification using pattern recognition with cepstral coefficients as features,” *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1933–1941, 1999.

[13] A. Eronen, “Comparison of features for musical instrument recognition,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, October 2001.

[14] P. Herrera, X. Amatriain, E. Batlle, and X. Serra, “Towards instrument segmentation for music content description: a critical review of instrument classification techniques,” in *International Symposium on Music Information Retrieval*, pp. 23–25, Plymouth, Mass, USA, October 2000.

[15] J. Marques and P. J. Moreno, “A study of musical instrument classification using gaussian mixture models and support vector machines,” Tech. Rep., Cambridge Research Laboratory, Cambridge, Mass, USA, June 1999.

[16] K. D. Martin, *Sound-source recognition: a theory and computational model*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass, USA, 1999.

[17] E. Wold, T. Blum, D. Keislar, and J. Wheaton, “Content-based classification, search, and retrieval of audio,” *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.

[18] J. Foote, “Automatic audio segmentation using a measure of audio novelty,” in *Proc. IEEE International Conference on*

Multimedia and Expo, vol. I, pp. 452–455, New York, NY, USA, August 2000.

[19] S. Pfeiffer, S. Fischer, and W. E. Effelsberg, "Automatic audio content analysis," in *Proc. ACM Multimedia*, pp. 21–30, Boston, Mass, USA, November 1996.

[20] T. Zhang and C.-C. Jay Kuo, Eds., *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*, Kluwer Academic Publishers, Boston, Mass, USA, February 2001.

[21] M. Casey, "General sound classification and similarity in MPEG-7," *Organized Sound*, vol. 6, no. 2, pp. 153–164, 2001.

[22] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method," in *Proc. ACM Multimedia*, pp. 203–211, Ottawa, Canada, October 2001.

[23] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. II, pp. 1331–1334, Munich, Germany, April 1997.

[24] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, NY, USA, 1996.

[25] G. Haus and E. Pollastri, "A multimodal framework for music inputs," in *Proc. ACM Multimedia*, pp. 382–384, Los Angeles, Calif, USA, November 2000.

[26] B. Flury, *A First Course in Multivariate Statistics*, Springer-Verlag, New York, NY, USA, 1997.

[27] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge, UK, 2000.

[28] F. Opolko and J. Wapnick, *McGill University Master Samples*, McGill Univeristy, Montreal, Quebec, Canada, 1987.

[29] C. Saunders, M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A. Smola, "Support vector machine reference manual," Tech. Rep., Royal Holloway Department of Computer Science Computer Learning Research Centre, University of London, Egham, London, UK, 1998, http://svm.dcs.rhbnc.ac.uk/.

[30] E. M. Hornbostel and C. Sachs, "Systematik der Musikinstrumente. ein Versuch," *Zeitschrift für Ethnologie*, vol. 46, no. 4-5, pp. 553–590, 1914, [English translation by A. Baines and K. P. Wachsmann, "Classification of musical instruments" *Galpin Society Journal*, vol. 14, pp. 3–29, 1961].

[31] I. Fujinaga and K. MacMillan, "Realtime recognition of orchestral instruments," in *Proc. International Computer Music Conference*, Berlin, Germany, August–September 2000.

[32] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New York, NY, USA, 1992.

[33] H. Späth, *Cluster Analysis Algorithms*, E. Horwood, Chichester, UK, 1980.

[34] A. Srinivasan, D. Sullivan, and I. Fujinaga, "Recognition of isolated instrument tones by conservatory students," in *Proc. International Conference on Music Perception and Cognition*, pp. 17–21, Sidney, Australia, July 2002.

Giulio Agostini received a "Laurea" in computer science and software engineering from the Politecnico di Milano, Italy, in February 2000. His thesis dissertation covered the automatic recognition of musical timbres through multivariate statistical analysis techniques. During the following years, he has continued to study the same subject and published his contributions to two IEEE international workshops devoted to multimedia signal processing. His other research interests are combinatorics and mathematical finance.

Maurizio Longari was born in 1973. In 1998, he received his M.S. degree in information technology from Università degli Studi di Milano, Milan, Italy, LIM (Laboratorio di Informatica Musicale). In January 2000, he started his research activity as a Ph.D. student at Dipartimento di Scienze dell'Informazione in the same university. His main research interests are symbolic musical representation, web/music applications, and multimedia database. He is a member of the IEEE SA Working Group on Music Application of XML.

Emanuele Pollastri received his M.S. degree in electrical engineering from Politecnico di Milano, Milan, Italy, in 1998. He is a Ph.D. candidate in computer science at Università degli Studi di Milano, Milan, Italy, where he is expected to graduate at the beginning of 2003 with a thesis entitled "Processing singing voice for music retrieval." His research interests include audio analysis, understanding and classification, digital signal processing, music retrieval, and music classification. He is cofounder of Erazero S.r.l., a leading Italian multimedia company. He worked as a software engineer for speech recognition applications at IBM Italia S.p.A. and he was a consultant for a number of companies in the field of professional audio equipments.