Content-Based Classification of Musical Instrument Timbres

Giulio Agostini Maurizio Longari Emanuele Pollastri

Laboratorio di Informatica Musicale - L.I.M.
Dipartimento di Scienze dell'Informazione
Università Statale degli Studi di Milano
Via Comelico 39
20135 Milano - Italy
agostini@lalim.lim.dsi.unimi.it, {longari, pollastri}@dsi.unimi.it

Abstract

A set of features extracted from audio sources is investigated for content-based classification of musical instrument timbres. The adopted features describe spectral characteristics of monophonic sounds and rely on the previous segmentation of the signal and the estimation of pitch. The dataset is composed by 1007 tones from 27 musical instruments ranging from orchestral sounds (strings, woodwinds, brass) to pop/electronic instruments (bass, electric and distorted guitar). The extracted features are then classified by widely used pattern recognition techniques. A thorough evaluation of the resulting performances and comparative analysis with previous works is presented. Quadratic Discriminant Analysis shows an error rate of 7.19% for the individual instruments and 3.23% for instrument families. These results are by far superior to the performances of other classification methods (Canonical Discriminant Analysis, Support Vector Machines, Nearest Neighbours). The use of a machine-built decision hierarchy did not improve the results.

1 Introduction

The introduction of languages for sound authoring, like CSound, or the more recent Structured Audio Orchestra Language (SAOL) in the newborn MPEG-4 standard, and languages devoted to describe audio content, like in the forthcoming MPEG-7 standard, revive the interest in automatic music understanding. A great number of commercial applications could soon be available for both entertainment and professional appliances, thus boosting research efforts in the multimedia scientific community.

An interesting application in the area of sound databases is the automatic classification of audio sources by musical instrument timbre, and this is the goal of the present work. Timbre differs from the other sound attributes, namely pitch, loudness, and duration, because it is ill-defined. 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" [1]. In other words, it is not possible to associate a physical quantity to the perceptual experience that we call "timbre."

In this paper, various classification methods have been employed over a set of features extracted from audio sources. The results will be compared to those reported in other works. Tests have been carried out with labelled sounds, i.e. using supervised classification. Issues about perceptual similarity have not been addressed; rather, our objective is the organization of sounds for multimedia libraries. An indexing schema of musical sounds should rely on a selection of audio descriptors that is reduced in number and significant. At the same time, a classification algorithm is needed in order to organize these descriptors into groups of similar timbres and to retrieve music information by content.

2 Related Work

A complete review of studies on timbre classification is out of our scope. For the interested reader, a recent paper has been presented by Serra et al. [8]. Previous works on musical instrument identification primarily focused either on feature extraction techniques or on classification methods, rarely on both. Researchers with a background on music signal analysis employed a wide range of features, justifying their choice in terms of musical relevance, brightness, spectral synchronicities, harmonicity, and so forth, but they used simple classification algorithms. On the other hand, works from other research areas used to simplify the feature extraction process in favour of more powerful classification techniques. For instance, in [5], 44 temporal and cepstral features are classified by means of a k-Nearest Neighbours algorithm and a Gaussian classifier. In

other studies, besides the introduction of advanced methods like Support Vector Machines [12] and Neural Networks [3, 11], a basic set of features have been extracted from audio (for instance: Mel Cepstrum Coefficients, Short-Time RMS-energy) or tests have been carried out with a limited amount of data (8 instruments or less).

As we mentioned earlier, the real-world applications envisioned by the automatic instrument identification span the domain of multimedia databases. There exist two implementations that allow searching sounds by similarity in digital archives: A commercial product by Musclefish called SoundFisher [18], and Studio Online, which is derived by researches conducted at IRCAM [9]. In the mid-long term, the early (possibly assisted) audio annotating systems should appear, such as an extractor of MPEG-7-like descriptors.

3 Feature Extraction

A great deal of work has been done to explore acoustic and perceptual features related to timbre. Since the first studies by Grey [7], it has been clear that we are dealing with a multi-dimensional attribute, which includes spectral and temporal features. An example of the former is the harmonic spectral centroid which corresponds to the perceived "brightness" of a sound, while the envelope attack time, which is bound to the "sharpness" of sounds, regards the latter.

A considerable number of features is currently available in the literature, each one describing some aspects of audio content. Since features are usually calculated out of a certain amount of samples, which is normally very small compared to the total duration of a tone, we must face the problem of summarizing their temporal evolution into a small set of values. Mean, standard deviation, skewness and auto-correlation have been the preferred strategies for their simplicity, but more advanced methods like Hidden Markov Models could be employed, as illustrated in [19]. 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 [4], which states: 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.

In this work, a set of features related to the harmonic properties of sounds is extracted from monophonic musical signals. The number of features implemented is small compared to previous works by Martin [13] and Klapuri [5]. The extraction of the descriptors relies on a number of preliminary steps, namely temporal segmentation of the signal, detection of the fundamental frequency and the estimation of the harmonic structure (Figure 1). The evaluation of automatic classification based only on spectral features is one of the main goals of our work. As we will show in Section 6, we achieved very satisfactory results without employing any temporal features.

3.1 Audio Segmentation

The aim of the first stage is the temporal segmentation of the audio signal 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. The output of this segmentation is a list of non-silent events (starting and ending points). A simple procedure based on energy evaluation is briefly described here. The signal is first processed with a band-pass Chebyshev filter of order five; cut-off 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 RMS-energy curve is computed with the same frame size. By comparing the energy to an absolute threshold, we find out a rough estimate of the boundaries of the events. A finer analysis is then performed at a 5 ms frame to determine actual on/off-sets; in particular we look for a 6 dB step near every rough estimate. This algorithm performs satisfactorily

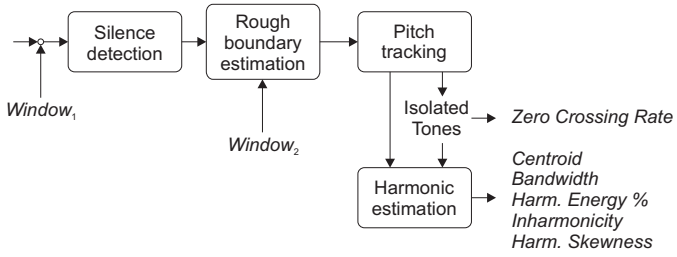


Figure 1: Block diagram of the feature extraction process.

for moderately noisy signals and isolated tones. In case of real executions, it may fail to detect some tone transitions, but the next step often fixes this problem.

3.2 Pitch Tracking

Pitch deserves a special place in our research, since it enables us to refine signal segmentation and it is the basic value for the calculation of some spectral features. Through pitch detection, we can identify notes that are not well defined by the energy curve or that are possibly played legato. At frame level, instantaneous values of the fundamental frequency are used to estimate features related to the harmonic structure. The pitch-tracking algorithm employed follows the one presented in [14], so it will not be described here. The output of the pitch tracking is the average value (in hertz) of each note hypothesis, a frame by frame value of pitch and a value of accuracy that measures the uncertainty of an estimate.

3.3 Calculation of Features

From each tone isolated through the procedure just described, a set of nine features is extracted frame by frame and their means and standard deviations are stored as descriptors for that event (Figure 2). Thus, we collect a total of 18 features for each tone. Pitch values (f_0) estimated in the previous stage are used only as a reference by the feature extraction algorithm. The signal is analysed with half-overlapping windows and smoothed with a Hamming function. The size of the analysis window is variable in order to have a frequency resolution of at least $1/24^{\text{th}}$ of octave, even for the lowest tones. Short-Time Fourier Analysis is then adopted for spectrum estimation.

Feature number (mean and standard deviation)	Feature name	Formula
1–2	Zero Crossing Rate	$z = \sum_n \big\lvert \operatorname{sgn}[s(n)] - \operatorname{sgn}[s(n-1)] \big\rvert / 2$ $\operatorname{sgn}(x) = \begin{cases} +1 & \text{if } x \geq 0, \\ -1 & \text{if } x < 0. \end{cases}$
3–4	Spectral Centroid	$c = \dfrac{\sum_{f=f_{\min}}^{f_{\max}} f \cdot E(f)}{\sum_{f=f_{\min}}^{f_{\max}} E(f)}$
5–6	Bandwidth	$b = \dfrac{\sum_{f=f_{\min}}^{f_{\max}} \lvert c - f \rvert E(f)}{\sum_{f=f_{\min}}^{f_{\max}} E(f)}$
7–14	Harmonic Energy Percentage	$E_{p_i} = \dfrac{\sum_{f=f_{\mathrm{L}_i}}^{f_{\mathrm{R}_i}} E(f)}{\sum_{f=f_{\min}}^{f_{\max}} E(f)}$ $f_{\mathrm{L}_i} = p_i - 1/24\,\text{oct}$ $f_{\mathrm{R}_i} = p_i + 1/24\,\text{oct}$ $1 \leq i \leq 4$
15–16	Inharmonicity	$\delta = \sum_{i=1}^{4} \dfrac{\lvert p_i - i \cdot f_0 \rvert}{i \cdot f_0}$
17–18	Harmonic Energy Skewness	$h = \sum_{i=1}^{4} \dfrac{\lvert p_i - i \cdot f_0 \rvert}{i \cdot f_0} E_{p_i}$

Figure 2: Description of the extracted features.

First, mean and standard deviation of *zero crossing rate* normalized with respect to the size of the window, *spectral centroid* (i.e. the centre of gravity of the spectrum) and *bandwidth* (or magnitude-weighted differences between the spectral components and the centroid) are calculated, see Figure 2. Then, the first four partials (p_i) are estimated as the most prominent peaks of the spectrum in a range of $1/12^{\text{th}}$ of octave, centred at frequencies f_0, $2f_0$, $3f_0$, and $4f_0$. We called the cumulative distance between the estimated partials and their theoretic value *inharmonicity*. Power spectral density of the first four bands centred at the partials and $1/12^{\text{th}}$ of octave wide are now normalized with respect to the total energy. In other words, we keep the percentage of total energy contained in each partial. Finally, we considered a novel feature (*harmonic energy skewness*), which is defined as the sum of the energy confined in the partial regions, multiplied by the respective inharmoncities.

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 to associate 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 [6] 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 $f_i(\cdot)$ of each class

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

where π_i is the *a priori* probability that an observation belongs to the i-th 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 i-th class every vector $\mathbf{y}$ for which

$$\pi_i f_i(\mathbf{y}) > \pi_j f_j(\mathbf{y}) \qquad \forall i \neq j. \tag{2}$$

Unfortunately, PDFs $f_i(\cdot)$ are generally unknown. Nonetheless, one 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

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

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{j=1}^{N} \mathbf{d}_{ij} \qquad \text{and} \qquad \mathbf{S}_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (\mathbf{d}_{ij} - \mathbf{m}_i)(\mathbf{d}_{ij} - \mathbf{m}_i)', \tag{3}$$

using the N_i observations $\mathbf{d}_{ij}$ available for the i-th 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, as we pointed out, this is the optimal classifier for normal mixtures, it could lead to sub-optimal error rates in practical cases, for two reasons. First, classes can depart sensibly from the assumption of normality. A subtler source of errors is the fact that with this method the actual distributions remain unknown, since we only have their 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 maximising 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, and for $c > 2$ classes, CDA does the same thing using a similar criterion.

Computationally equivalent to QDA, 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 "very different," QDA has lower error rates. QDA is also to be preferred in presence of long tails and pronounced kurtosis, whereas a moderate skewness suggests to use CDA.

4.2 k-Nearest Neighbours

This is one of the most popular non-parametric technique 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).

4.3 Support Vector Machines

The Support Vector Machines (SVM) are a recently developed approach to the learning problem [2]. The aim is to find the linear hyperplane that best separates observations belonging to different classes.

Suppose we have a set of linearly separable training samples $\mathbf{d}_1, \ldots, \mathbf{d}_N$, 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 i-th 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 $\mathbf{w}$ subject to class separation constraints. The optimal solution can be viewed in a dual form by applying the Lagrange Theory and imposing the conditions of stationariness. The Support Vectors are defined as the input samples $\mathbf{d}_i$ for which the respective Lagrange multiplier is non-zero, 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.

For the non-linearly separable case, the samples are projected through a non linear function $\Phi(\cdot)$ from the input space Y in a higher-dimensional space (the transformed space[1] T). Since the high number of dimensions increases the computational effort, it is possible to introduce the *kernel functions*

$$K(\mathbf{y}, \mathbf{z}) = \langle \Phi(\mathbf{y}) \cdot \Phi(\mathbf{z}) \rangle \,, \tag{4}$$

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 best linear hyperplane can always be found, but, in practice, a solution can be heuristically obtained. Obviously, 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 (RBFs) and polynomial kernels are widely used in image recognition, speech recognition, hand-written digit recognition, and protein homology detection problems.

5 Experiment

An extended collection of musical instruments tones is essential for training and testing classifiers. To achieve results comparable to the previous works by Martin [13] and Klapuri [5], our dataset comes from the MUMS (McGill University Master Samples) CDs [15], which are a library of isolated sample tones from a wide number of musical instruments, played with several articulation styles and covering the entire pitch range. A large dataset is needed, 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 Section 3, thus 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, even though the total number of musical instruments considered was 27, the classification results reported in the next section can be claimed to hold for a set of 30 instruments.

MUMS CDs provided standard Audio CD quality files—sampling frequency of 44.1 kHz, 16 bit dynamic resolution—which have been analysed by the feature extraction algorithms. If the accuracy of a pitch estimate is below a pre-defined threshold, the corresponding tone is rejected from the training set. Following this procedure, the number of tones accepted for training/testing was 1007 in total. We adopted a leave-one-out error rate estimation method for each of the classifiers tested: CDA, QDA, k-NN, k-NN with kernel (i.e. the

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

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

input space is modified according to a kernel function) and SVM. Tests have been carried out with a growing number of classes (13, 17, 20, and 27 instruments), and classifiers that clearly performed unsatisfactorily with a smaller set of instruments have not been employed in the subsequent experiments. k-NN has been tested with $k = 1, 3, 5, 7$ and with 3 different distance metrics (1-norm, Euclidean 2-norm, 3-norm). For SVM, we adopted a software tool developed at the Royal Holloway University of London [16]. Input values have been normalized independently and we chose a multi-class classification method that trains $c(c-1)/2$ binary classifiers, where c is the number of instruments.

6 Results

For each experiment, results have been evaluated by means of confusion matrices and overall success rates. Although we put the emphasis at the instrument level, we have also grouped instruments belonging to the same family (strings, brass, woodwinds and the like), extending Sachs' taxonomy [10] with the inclusion of "rock strings" (deep bass, electric guitar, distorted guitar). The SVM classifier has been tested with a subset of 17 and 20 musical instruments and with various kernels in order to explore their performances. Since RBF kernels obtained the best results, this SVM classifier has been chosen for the classification of 27 instruments. k-NN did not present a consistent trend, going from 13 to 27 instruments, except that 1-NN with 1-norm distance always performed better than 3/5/7-NN in combination with the other distance metrics. The introduction of kernel did not improve the error rate; for instance, 1-NN performed with 71% success rate on 20 instruments with polynomial kernel of order 1 and 74% with no kernel.

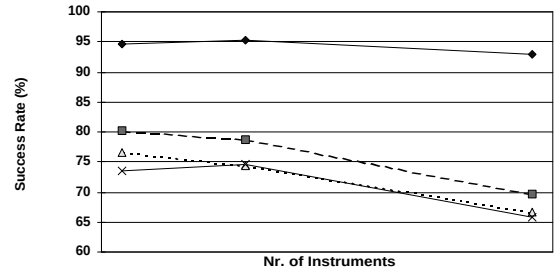
Figure 3 provides a graphical representation of the best results at the instrument level, achieved with a dataset of 17, 20 and 27 instruments. QDA performed better than the other classifiers in every test, with an impressive success rate of 92.81% for 27 instruments and with an almost stable trend (from 94.7% to 92.81%). The confusion matrix relative to this case is depicted in Table 1. Most of the misclassifications are within the correct instrument family (e.g. doublebass classified as cello), except for piano and cello, classified respectively as viola pizzicato (13% of piano tones) and classic guitar (15% of cello tones). Comparing the QDA confusion matrices for 13 and 27 instruments, it is remarkable that success rates for the instruments in common are the same. CDA and 1-NN have never obtained momentous results, in fact success rates range from 65.74% (k-NN, 27 instruments) to 76.63% (CDA, 17 instruments).

SVM achieved the second best score, showing a plunge as the number of instruments increases (from 80.20% to 69.71%). If we compare our results with the ones reported by Marques [12] (30% error rate with 8 instruments), the SVM classifiers presented here had an error rate of 20%, despite our classes are twice as much. This can be only partially explained by the different training sets employed, so we draw the conclusion that our set of features is better suited for describing musical timbres than the one employed by Marques [12], which is derived from the speech-recognition area.

At the instrument family level, classification results based on 27 instruments are shown in Table 2. Our best success rate (96.77%) was better than any other results we are aware of, although the different taxonomy employed by Klapuri and the introduction of new families with respect to Martin makes a direct comparison difficult. The identification of a broader group of instruments, namely "pizzicati" and "sustained," was achieved with an average success rate of 97.25% that is lower than those reported by Martin and Klapuri (99%). This was to be expected, for two main reasons. First, we did not introduce any feature related to the time envelope of sounds. For instance, cello bowed is classified as sustained in 82% of trials and as pizzicato in 17% (confusion with classic guitar and viola pizzicato). Also, the family of pizzicati in our dataset is larger than the ones in cited experiments since it includes piano, harpsichord, harp and classic guitar.

Although k-NN was one of the favourite techniques in previous works on timbre classification, it must be noticed that it showed the worst performance with success rates similar to those reported by Martin [13]. Furthermore, the change in the extracted features did not affect the performance. Using the best features for each instrument and k-NN, Klapuri reported an 80% success rate, which is very far from QDA performances for a comparable dataset. Moreover, unlike Klapuri, we did not consider pitch ranges in our classifications.

In one of our experiments, we have also made use of a decisional tree. Instead of imposing the structure, though, as Martin and Klapuri did, we used a hierarchical clustering algorithm [17], because we thought that imposing a hierarchy rigidly based on the traditional taxonomy of western instruments could have had a negative impact on the results. Even with this machine-built hierarchy, the classification of 27 instruments, using CDA in each decisional node, brought the results down to 59.89% (against 66.74% with flat CDA classification). With this preliminar experient, we thus confirm Klapuri's conclusions, that hierarchical classification does not improve the error rate.



	17	20	27
QDA	94.70	95.34	92.81
SVM	80.20	78.64	69.71
CDA	76.63	74.41	66.74
KNN	73.51	74.54	65.74

Figure 3: Graph showing classifiers performances for different number of instruments.

Stimulus input / Recognised as	Hamburg Steinway	Harpsichord	Classic Guitar	Harp	Deep Electric Bass	Deep Elect. Bass Slap	Electric Guitar	Distorted Elect. Guitar	Violin Pizz.	Viola Pizz.	Cello Pizz.	Doublebass Pizz.	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
Hamburg Steinway	73																										
Harpsichord		100																									
Classic Guitar	8		89							4					15	5											
Harp				94																							
Deep Electric Bass				3	100																						
Deep Elect. Bass Slap				3		100																					
Electric Guitar							100																				
Distorted Elect. Guitar								100																			
Violin Pizz.			2						85																		
Viola Pizz.	12		7						15	96	11	4			2	4											
Cello Pizz.											86	8															
Doublebass Pizz.												88															
Violin Bowed													82	5													
Viola Bowed													18	95										2.5			
Cello Bowed	2		2												80	19											
Doublebass Bowed	5														3	72											
Flute																	100										
B. Plenum Organ																		100									
Accordion																			100								
Bassoon																				100							
Oboe																					97			2.5			
English Horn																						97					
Eb Clarinet																					3	3	100				
Sax																								95			
C Trumpet																									100		
French Horn																										100	
Tuba										3																	100
Family success (%)	89.72				100				98.15				93.11				99.66								100		
Pizz./Sust. success (%)	95.14												97.70														

Table 1: Confusion matrix for $c = 27$ instruments, classified with a flat QDA classifier.

Classifier	Family Success Rate (%)	Pizzicato/Sustained Success Rate (%)
QDA	96.77	97.25
CDA	79.10	79.27
SVM	78.04	78.58
k-NN	76.61	77.47

Table 2: Summary table for higher levels of abstraction, with $c = 27$ instruments.

7 Discussion and Further Work

It has been demonstrated that broadly used classifiers could not provide comparable results to QDA performances. Since QDA is the optimal classifier under multivariate normality hypotheses, the results seem to suggest that the features we extracted from isolated tones follow such distribution. To validate this hypothesis a series of statistical tests is undergoing on the dataset. Although hierarchical classification could lead to faster and more flexible classifiers (e.g. selection of the best features or the best classification method in each decisional node), with these early results we found that it is of no advantage. Our feature set still lacks of temporal descriptors of the signal, as it has been made clear by the poor pizzicato/sustained discrimination. Thus, we plan to introduce features like log attack slope or, more audaciously, new timing cue schemes like the cited HMMs. The introduction of new features will be gradually accomplished since the compactness of the representation is one of the requirements for efficient database architectures. A new session of tests with music samples extended to percussive sounds and with live-recorded musical instruments has already started.

References

[1] American National Standards Institute. *American National Psychoacoustical Terminology.* S3.20. American Standards Association, New York, 1973.

[2] N. Cristianini, J. Shawe-Taylor. "Support Vector Machines and other kernel-based learning methods." Cambridge University Press, 2000.

[3] P. Cosi, G. De Poli, P. Prandoni. "Timbre characterization with Mel-Cepstrum and neural nets." Proceedings of the ICMC 1994, 42–45, 1994.

[4] L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag,1996.

[5] A. Eronen, A. Klapuri. "Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features." IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2000.

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

[7] J. M. Grey. "Multidimensional perceptual scaling of musical timbres." *Journal of the Acoustical Society of America* **61**(5), 1270–1277, 1977.

[8] P. Herrera, X. Amatrian, E. Batlle, X. Serra. "Towards instrument segmentation for music content description: a critical review of instrument classification techniques." International Symposium on Music Information Retrieval, Plymouth (MA), 23–25 October, 2000.

[9] P. Herrera, S. McAdams, G. Peeters. "Instrument sound description in the context of MPEG-7." Proceedings of the ICMC 2000, Berlin, Germany, 27 August–1 September, 2000.

[10] E. M. Hornbostel, C. Sachs. "Systematik der Musikinstrumente. Ein Versuch." Zeitschrift für Ethnologie, **46**, 1914. (English translation by A. Baines and K. P. Wachsmann. Galpin Society Journal, **14**, 1961.)

[11] I. Kaminskyj, A. Materka. "Automatic source identification of monophonic musical instrument sounds." Proceedings of the 1995 IEEE International Conference on Neural Networks, 189–194, 1995.

[12] J. Marques, P. J. Moreno. "A study of musical instrument classification using Gaussian Mixture Models and Support Vector Machines." Tech.Report 99-4, Compaq Cambridge Research Laboratory, 1999.

[13] K. D. Martin. *Sound-Source Recognition: A Theory and Computational Model.* Ph.D. Thesis, Massachussets Institute of Technology, 1999.

[14] E. Pollastri. "Melody retrieval based on approximate String-Matching and Pitch-Tracking Methods." Proc. of XII Colloquium on Musical Informatics, Gorizia, 151–154, Oct. 1998.

[15] F. Opolko, J. Wapnick. "McGill University Master Samples." McGill University, Montreal, 1987.

[16] C. Saunders, M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, A. Smola. "Support Vector Machine reference manual." Royal Holloway Department of Computer Science Computer Learning Research Centre.

[17] H. Späth. *Cluster Analysis Algorithms.* Ellis Horwood Ltd., Chichester, 1980.

[18] E. Wold, T. Blum, D. Keislar, J. Wheaton. "Content-based classification, search, and retrieval of audio." IEEE Multimedia, 27–36, Fall 1996.

[19] T. Zhang, C. C. J. Kuo. "Hierarchical classification of audio data for archiving and retrieving." IEEE ICASSP, **6**, 3001–3004, Phoenix, March 1999.