MUSICAL INSTRUMENT TIMBRES CLASSIFICATION WITH SPECTRAL FEATURES

G. Agostini **M. Longari** **E. Pollastri**
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 - In this work, a set of features is evaluated for musical instrument recognition 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 and without employing any hierarchical structure, Quadratic Discriminant Analysis shows the lowest error rate (7.19% for the individual instrument and 3.13% for instrument families), outperforming all the other classification methods (Canonical Discriminant Analysis, Support Vector Machines, Nearest Neighbours). The most relevant features are demonstrated to be the inharmonicity, the spectral centroid and the energy contained in the first partial.

INTRODUCTION

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

Focussing on this area, the forthcoming MPEG-7 standard should provide a list of metadata for multimedia content, nevertheless two important aspects still need to be explored further. First, it is needed to identify what are the best features suited for a particular task. 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 present work with the objective of automatic timbres classification for sound databases.

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 to a particular physical quantity. The uncertainty about the notion of timbre is reflected by the huge amount of studies that have tackled this problem. Early works on timbre recognition focussed on the exploration of possible relationship between the perceptual and the acoustic domain [3]. Recently, the diffusion of multimedia databases has brought to the fore the problem of musical instrument identification out of a fragment of audio signal and a recent overview on the topic is presented in [4]. Usually, this task is accomplished by training the classifier with labelled instances (supervised learning).

FEATURE EXTRACTION

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 the results section, a set of features related mainly to the harmonic properties of sounds allows a simplified representation of data, without losing important characteristics. Moreover, reducing the number of features prevents from incurring into the so called *curse of dimensionality* [1]. The extraction of descriptors relies on a number of preliminary steps, that is temporal segmentation of the signal, detection of the fundamental frequency and the estimation of the harmonic structure (Figure 1).

After band-pass filtering the signal, a procedure based on energy evaluation is carried out in order to have a rough estimation of event boundaries. A RMS-energy curve is computed on the windowed signal (Hamming, 46 ms) and compared with an absolute threshold (silence detection). A finer analysis is then conducted at a 5 ms frame rate to 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*. The pitch-tracking algorithm employed follows the one presented in [8], 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 an accuracy value that measures the uncertainty of the estimate.

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 a tone. The Zero Crossing Rate is measured

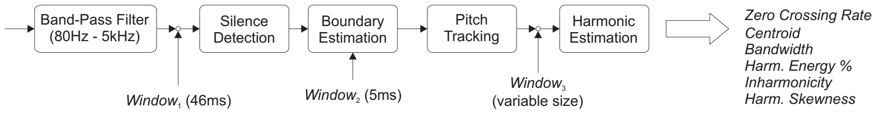


Figure 1: Description of the feature extraction process (see text for details).

Pizzicati			Sustained		
Piano et al.	Rock Strings	Pizz. Strings	Strings	Woodwinds	Brass
Piano Harpsichord Classic Guitar Harp	Electric Bass Elect. Bass Slap Electric Guitar Dist. Elect. Guitar	Violin pizzicato Viola pizzicato Cello pizzicato Doublebass pizz.	Violin bowed Viola bowed Cello bowed Doublebass bowed	Flute Organ Accordion Bassoon Oboe English Horn E♭ Clarinet Sax	C Trumpet French Horn Tuba

Figure 2: Taxonomy of the instruments employed in the experiments.

directly from the waveform and normalized with respect to the size of the window. Then, the harmonic structure of the signal is evaluated through 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 $1/24^{\text{th}}$ of octave, even for the lowest tones ($1024 \div 8192$ samples). From the harmonic analysis we calculate the remaining features: spectral centroid (i.e. the centre of gravity of the spectrum), bandwidth (or magnitude-weighted differences between the spectral components and the centroid), inharmonicity (cumulative distance between the estimated partials and their theoretic values), percentage of energy contained into the first four partials, and harmonic energy skewness (sum of energy confined in the partials region, multiplied by the respective inharmoncities).

EXPERIMENT

The dataset adopted has been extracted by the MUMS (McGill University Master Samples) CDs [9], 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). Since the saxophones (alto, soprano, tenor, baritone) have been collapsed to a single instrument class, the total number of instruments in our tests will be 27 (Figure 2). The audio files 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 is 1007 in total. Various classification techniques have been implemented and tested: Canonical Discriminant Analysis (CDA), Quadratic Discriminant Analysis (QDA), Nearest Neighbours (k-NN) and Support Vector Machines (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 [10]. A number of kernel functions has been considered (dot product, simple polyno-

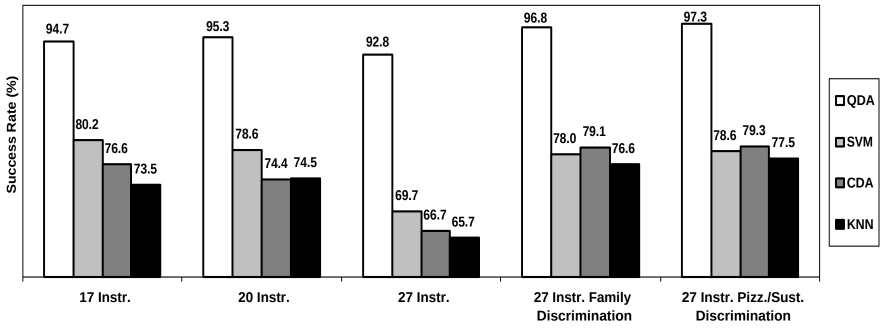


Figure 3: Graphical representation of the success rates for each experiment.

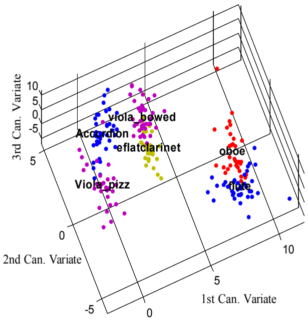
mial, Radial Basis Functions, linear splines, regularized Fourier). All error rates estimates reported in the next section have been computed using a leave-one-out procedure.

RESULTS

Figure 3 provides a graphical representation of the best results both at instrument level (17, 20 and 27 instruments) and at family level (pizzicato-sustained, instrument family). QDA performed better than any other classifier 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 second best score was achieved by SVM with Radial Basis Function kernel; it must be noted that increasing the number of instruments, SVM success rates decreased from 80.20% (17 instruments) to 69.71% (27 instruments). In comparison with the work by Marques and Cano [5], 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. CDA and k-NN never obtained momentous results (respectively 66.74% and 65.74% with 27 instruments). Among the k-NN classifiers, 1-NN with 1-norm distance metric obtained the best performance. Using a kernel function to modify the input space did not bring any advantage (71% with kernel and 74% without kernel for 20 instruments). A deeper analysis of the results achieved with QDA showed that 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). We have also calculated a list of the most relevant features through the forward selection procedure detailed in [7]. The values reported are the normalized versions of the statistics on which the procedure is based. They can be not strictly decreasing, because a feature might bring more information only jointly with other 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) stddandard deviation	0.116544

(a)



(b)

Figure 4: Most discriminating features for 27 instruments (a) and dataset for 6 instruments projected to the first three canonical variates (b).

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 Figure 4(a)). In Figure 4(b), the dataset is projected in the first three canonical variates, for a subset of 6 instruments. At the instrument family level, our best success rate (96.77%) was better than any other work we are aware of, although the different taxonomy employed by Klapuri [2] and the introduction of new families with respect to Martin [6] makes a direct comparison difficult. Pizzicato and sustained instruments were recognized with 97.24% success rate which is lower than those reported by Martin and Klapuri (99%) [6, 2]. However, the family of pizzicati in our dataset is larger than the ones in cited experiments. Moreover, we did not make use of any temporal feature.

In one of our experiments, we have also introduced a machine-built decisional tree. We used a hierarchical clustering algorithm [11] to build the structure. CDA or QDA methods have been employed at each node of the hierarchy. Even with this techniques, though, we could not improve the error rates; for instance, 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). As a final remark, the computational complexity of CDA and QDA are equivalent since they are both in the order of $\Theta(kp^2)$, p being the number of features and k the number of classes.

DISCUSSION AND FURTHER WORK

As it was demonstrated, a set of spectral features combined with QDA classifiers showed the best performances; other broadly used classifiers could not provide comparable results. 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 from pitch estimation.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Haus for his support and all the staff at L.I.M. (Laboratorio di Informatica Musicale). Special thanks to Prof. N. Cesa Bianchi for the helpful discussions about SVMS.

References

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

[2] A. Eronen and A. Klapuri, "Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features." in IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2000, Istanbul, June 2000.

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

[4] 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.

[5] J. Marques and 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.

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

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

[8] G. Haus and E. Pollastri, "A Multimodal Framework for Music Inputs" Proc. of ACM Multimedia 2000, Los Angeles, Nov. 2000.

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

[10] http://svm.dcs.rhbnc.ac.uk/

[11] H. Späth, *Cluster Analysis Algorithms.* E. Horwood Ltd., Chichester, 1980.