Musical Instrument Recognition by Pairwise Classification Strategies

Slim Essid, Gaël Richard, *Member, IEEE*, and Bertrand David

***Abstract*—Musical instrument recognition is an important aspect of music information retrieval. In this paper, statistical pattern recognition techniques are utilized to tackle the problem in the context of solo musical phrases. Ten instrument classes from different instrument families are considered. A large sound database is collected from excerpts of musical phrases acquired from commercial recordings translating different instrument instances, performers, and recording conditions. More than 150 signal processing features are studied including new descriptors. Two feature selection techniques, inertia ratio maximization with feature space projection and genetic algorithms are considered in a class pairwise manner whereby the most relevant features are fetched for each instrument pair. For the classification task, experimental results are provided using Gaussian mixture models (GMMs) and support vector machines (SVMs). It is shown that higher recognition rates can be reached with pairwise optimized subsets of features in association with SVM classification using a radial basis function kernel.**

***Index Terms*—Feature selection, Gaussian mixture model (GMM), genetic algorithms, inertia ratio maximization with feature space projection (IRMFSP), musical instrument recognition, pairwise classification, support vector machine (SVM).**

I. Introduction

THE NEED for multimedia content description has become a major issue as larger and larger digital data has been made available for millions of both amateur and professional end-users. This has been particularly scoped out by the light of MPEG-7 standardization effort [1]. As far as musical content is concerned, it is desired to obtain score-like representations at a high level of description, which implies the ability to extract characteristics such as genre, rhythm, melody, playing instruments, etc. One could then setup systems capable of executing requests such as "*find Hard-bop Sax solo played in C# in database*." Thus, musical instrument recognition capability stands as a key feature of such systems. Knowing the instruments involved in a given musical piece is in itself a useful information; but furthermore, it may help discover other musical characteristics such as genre (a piano, double bass and drums trio is likely to be a jazz trio) or played notes (multipitch detection or source separation could be easier knowing the playing instruments).

However, identifying instruments from complex mixtures involving more than one playing at a time remains a very difficult problem that has been addressed in a very few studies [2]–[6] with often important restrictions regarding the musical content with respect to instruments involved and played notes. Of course, such a goal is far more challenging, yet it is believed that a great deal of work still has to be carried out in the so-called monophonic or solo context wherein only one instrument is played at a time. In fact, it is considered as an essential effort in providing insights into musical instrument timbre and a basis for handling real world polyphonic music as it may be conducted under the most realistic conditions by using sound material excerpted from commercial recordings. Indeed, directions have been proposed to extend the processing from mono-instrument to poly-instrument content either by means of prior musical source separation (see [7] for example) or adapted classification strategies [8].

While describing the timbre of musical instruments has received early concern, especially in the musical acoustics and psychoacoustics community [9]–[13], machine recognition of musical instruments is a quite recent research area which came into act in the last decade. The majority of studies handled the problem using sound sources consisting of isolated notes [14]–[22]. There are two main advantages in such approaches. First, the simplification of signal processing stages concerned with feature extraction, hence the ability to use more sophisticated descriptors which are difficult to measure in the multinote case (see Section II). Second, several public sound databases of isolated notes are available and can easily be used for such studies [23]–[26]. However, adopting these conditions imply the loss of note-to-note transition information which is known to be a particularly important aspect of timbre. Moreover, it is still not very clear how to bring such work to useful user applications since it is not practicable, given the current state of the art, to proceed to note segmentation prior to instrument recognition; except for percussive instruments [27].

Fewer studies dealt with musical phrases from real solo performances [14], [28]–[36]. Much effort was primarily dedicated to propose relevant features for musical instrument recognition including temporal, spectral, and cepstral features as well as their variation and statistics over a certain time or frequency horizon. The effect of combining features was studied [30], [33], [37], and feature selection techniques were considered (for example, context dependent feature selection in a hierarchical classification scheme in [14], backward and sequential feature generation in [19], or recursive selection based on inertia ratio maximization in [21] and [38]).

Various popular classification strategies were also studied [39]. K-nearest neighborhood (KNN) algorithms were largely used in early work on isolated notes [14], [19], [40]–[42]. Discriminant analysis was used as preprocessing in [14] and

Manuscript received August 13, 2004; revised May 17, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Davies.

The authors are with the LTCI-CNRS, GET-Télécom Paris, 75014 Paris, France (e-mail: Slim.Essid@enst.fr; Gael.Richard@enst.fr; Bertrand.David@enst.fr).

Digital Object Identifier 10.1109/TSA.2005.860842

for classification in [42]. In [21], hierarchical Gaussian classifiers were exploited after a Box–Cox transformation had been applied to each feature. Neural networks were also examined in a number of studies (see [43] for example). Also, multivariate Gaussian models, Gaussian mixture models (GMMs), and hidden Markov models (HMMs) were considered (see [19], [20], [44], and [45] for example). For recognition on solo phrases, GMM [29], [30], [32], HMM [31], and support vector machines (SVMs) [32], [33], [46] were found successful.

In this paper, the focus is put on musical instrument recognition on solo (unaccompanied) performance. All effort is employed to enhance the different parts of the recognition system, and our main contributions are linked to the following.

- *The sound database:* a much larger and more varied sound database with respect to instrument instances, recording conditions and players is used (compared to related studies).
- *The features:* a wide selection of features is considered, including new proposals, and their efficiency studied through feature selection techniques, namely inertia ratio maximization and genetic algorithms.
- *The classification schemes:* both GMM and SVM are considered. For GMM, model orders are assessed with a Bayesian information criterion (BIC). As for SVM, different types of kernels are considered and their relative performance discussed. Moreover, the influence of the number of consecutive temporal observations to be used for decision is studied.

Another contribution is that we argue that it is advantageous to address the task of instrument recognition using a pairwise classification (one versus one) strategy. We show, through experimental work, that performing instrument pairwise feature selection and classification results in better recognition accuracy and enables better understanding of timbral differences.

The outline of the paper is the following. In Section II, we give an overview of the feature set considered for classification. Then, the feature selection algorithms used in this work are presented in Section III. Following a concise description of the theoretical background related to GMM, SVM, and classification by pairwise coupling (Section IV), we proceed to the experimental studies to assess the efficiency of our recognition system (Section V). Finally, we suggest some conclusions in Section VI.

II. Feature Extraction

Finding appropriate features to model the timbre of musical instruments has received much concern toward obtaining a representation of humans' perception of musical sound [13], [47]. Our approach is more pattern-recognition oriented, in the sense that we examine an important number of low-level features to be automatically processed by a feature selection algorithm in order to fetch the most efficient in discriminating the musical instruments. Clearly, it can be then difficult to interpret some of the low-level features obtained in terms of timbre modeling.

In marked contrast to other pattern recognition tasks such as speaker identification, there has been no real consensus in choosing a set of features amenable to successful instrument recognition. Several studies show that MFCC alone turn out to be inefficient for discriminating between certain instrument classes (see [33] for example). In fact, many other features have been proposed [14], [19], [39], [48] describing various sound qualities. Also, automatic generation of high-level music descriptors using genetic programming was attempted [49]. A number of these features become quite difficult to extract when dealing with musical phrases. Typically, note attack characteristics are not straightforward to evaluate since onset detection is already intricate in our case.[1] Thus, a set of features which can be extracted in a quite simple and robust manner was chosen. In the following, we present a brief description of the features used. All of them are extracted on a frame basis.

A. Classical Features

Temporal: They consist of the following:

- autocorrelation coefficients (AC), which represent the overall trend of the spectrum [48], they were reported to be useful in [51];
- zero crossing rates (ZCR), which are useful for discriminating periodic signals (small ZCR values) from noisy signals (high ZCR values).

Cepstral: Mel-frequency cepstral coefficients (MFCCs) are considered as well as their time first and second derivatives which are estimated over a number of successive frames [52].

Spectral: These include a subset of features obtained from the statistical moments, namely the spectral centroid (Sc), the spectral width (Sw), the spectral asymmetry (Sa) defined from the spectral skewness, and the spectral flatness (Sf) defined from the spectral kurtosis. These features have proven to be successful for drum loop transcription [27] and for musical instrument recognition [33]. They are denoted by $\mathrm{Sx} = \{\mathrm{Sc}, \mathrm{Sw}, \mathrm{Sa}, \mathrm{Sf}\}$. Their time derivatives ($\delta\mathrm{Sx}$) are also computed in order to provide an insight into spectral shape variation over time. It is worth to note that $\delta\mathrm{Sc}$ can be seen as a quality of the vibrato playing technique since it embeds some frequency modulation information [19]. A more precise description of the spectrum flatness is also used, namely MPEG-7 audio spectrum flatness (ASF) [1] which is processed over a number of frequency bands. Indeed, this feature subset was found to be very useful for our task [33]. Moreover, frequency derivative of the constant-Q coefficients (describing spectral "irregularity" or "smoothness") are extracted as they were reported to be successful by Brown [30]. Another useful feature consisted in a measure of the audio signal frequency cutoff (Fc) (also called frequency rolloff in some studies [48]). It is computed as the frequency below which 99% of the total spectrum energy is accounted.

Amplitude Modulation Features (AM): These features are meant to describe the "tremolo" when measured in the frequency range 4–8 Hz, and the "graininess" or "roughness" of the played notes if the focus is put in the range 10–40 Hz [19]. First, temporal amplitude envelopes are computed using a low-pass filtering (10-ms half Hanning window) of signal

[1] note that onset detection for a differentiated transient/steady processing in the recognition process is tractable at the cost of additional complexity in the signal processing and decision stages, see [50] for further details

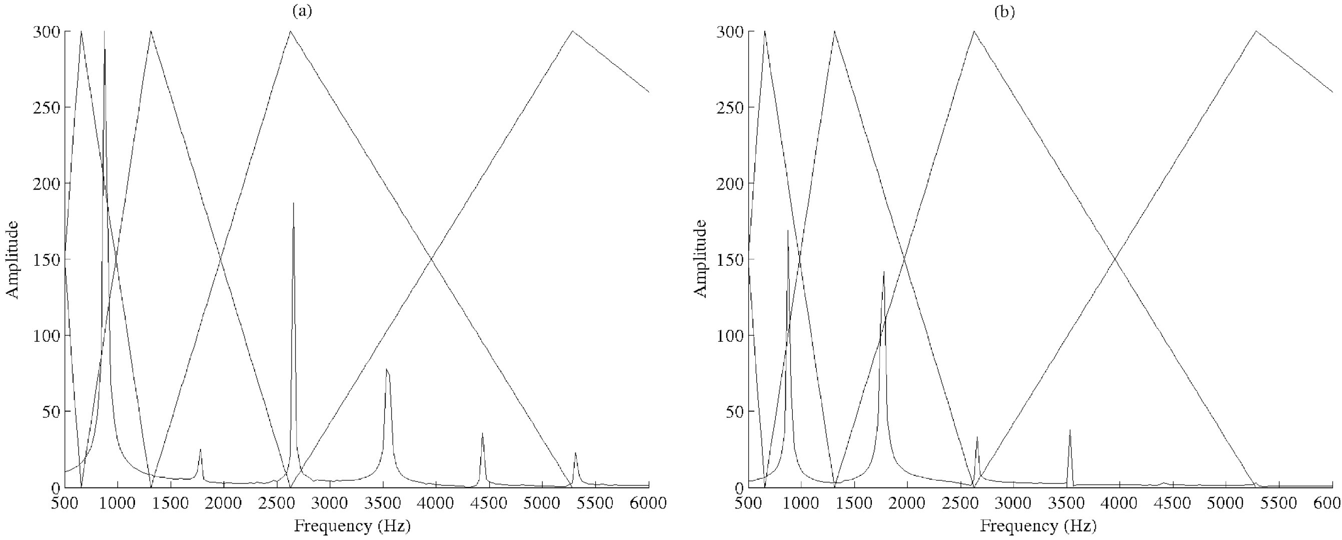


Fig. 1. Amplitude spectrums of (a) alto sax and (b) Bb clarinet playing the same note A4 and the octave band filterbank. In the second subband, higher OBSI will be measured for the Bb clarinet; in the third and forth subbands, higher OBSI for the alto sax.

absolute complex envelopes, then a set of six coefficients is extracted as described in Eronen's work [19], namely, AM frequency, AM strength and AM heuristic strength (for the two frequency ranges). Two coefficients are appended to the previous to cope with the fact that an AM frequency is measured systematically (even when there is no actual modulation in the signal); they are the product of tremolo frequency and tremolo strength, as well as the product of graininess frequency and graininess strength.

B. New Features

Octave Band Signal Intensities (OBSI): The idea behind this new feature set is to capture in a rough manner the power distribution of the different harmonics of a musical sound without recurring to pitch-detection techniques. In fact, a precise measure of frequencies and amplitudes of the different partials is not required for our task. One rather needs to represent the differences in harmonic structure between instruments. This can be achieved by considering a proper filterbank, designed in such a way that the energy captured in each subband vary for two instruments presenting different energy distribution of partials. Thus, we consider an octave band filterbank with triangular frequency responses. Filter edges are mapped to musical note frequencies starting from the lowest piano note A1 (27.5 Hz). For each octave subband, the maximum of the frequency response is reached in the middle of the octave subband. Important overlap is kept between adjacent channels (half octave). We then measure the log energy of each subband (OBSI) and the logarithm of the energy ratio of each subband sb to the previous $sb - 1$ (OBSIR).

As a result, the energy captured in each octave band as well as the energy ratio of one band to the previous will vary for two instruments having different harmonic structures. Additionally, in most cases, coarse locating of the fundamental frequency (f_0) is achieved since its octave range can be deduced from the first peak in the OBSI function. Fig. 1(b) gives an illustration of this discussion with alto sax and Bb clarinet playing the same musical note A4. For example, one can easily observe that the Bb clarinet has more energy in the second subband appearing in the plot than the alto sax, while the atlo sax has more energy than the Bb clarinet in the third and forth subbands. In fact, it is known that the Bb clarinet is characterized by the prominence of its odd harmonics and OBSI/OBSIR attributes allow us to describe such a characteristic.

III. Feature Selection Techniques

In many classification tasks, a very high number of potentially useful features can be considered. Often, some of these features are "noisy" or redundant with others. Though it is sometimes practicable to use all features for classification, it is clearly suboptimal to do so, especially if comparable performance can be achieved using a reduced set of features. Consequently, feature selection or transformation techniques are classically utilized both to reduce the complexity of the problem (by reducing its dimensionality) and to retain only the information that is relevant in discriminating the possible classes.

Feature transform techniques [typically principal component analysis (PCA) [53]] present the inconvenience of requiring that all candidate features be extracted at the stage of test (before the transform found during training is applied to them). Additionally, the transformed features are difficult to interpret, which is a major drawback if one expects to gain some understanding of the classes (here related to musical timbre).

Therefore, feature selection is often preferred to feature transformation, both to avoid extracting irrelevant features during testing and to be able to exploit the resulting descriptors in an intuitive way. By feature selection (FS), a subset of d features is selected from a larger set of D candidates. The selected subset is required to include the most relevant features, i.e., the combination yielding the best classification performance. Several strategies have been proposed by the statistical machine learning community [54]–[56] to tackle the problem. They can be classified into two major categories: the "filter" algorithms use the initial set of features intrinsically, whereas the "wrapper" algorithms relate the FSA to the performance of the classifiers to be used. The latter are more efficient than the former, but more complex. In this paper, we choose to exploit approaches that were proposed in previous work on musical instrument recognition, namely genetic algorithms (GAs) [41] and inertia ratio maximization using feature space projection (IRMFSP) [21], [34]. The efficiency of GA for feature selection has been argued in several studies [57]–[62]. IRMFSP present the advantage of being a simple and intuitive approach.

In the following, we present an overview of the IRMFSP algorithm and GAs. The particularity of our approach is to proceed to class pairwise feature selection (see Section III-C).

A. IRMFSP

Feature selection is made iteratively with the aim to derive an optimal subset of d features among D, the total number of features. At each step i, a subset $\mathbf{X}_i$ of i features is built by appending an additional feature to the previously selected subset $\mathbf{X}_{i-1}$. Let K be the number of classes, N_k the number of feature vectors accounting for the training data from class k, and N the total number of feature vectors ($N = \sum_{k=1}^{K} N_k$).

Let $\mathbf{x}_{i,n_k}$ be the n_kth feature vector (of dimension i) from class k, $\mathbf{m}_{i,k}$ and $\mathbf{m}_i$ be, respectively, the mean of the vectors of the class k $(\mathbf{x}_{i,n_k})_{1\leq n_k \leq N_k}$ and the mean of all training vectors $(\mathbf{x}_{i,n_k})_{1\leq n_k\leq N_k;\, 1\leq k\leq K}$.

Features are selected based on the ratio r_i (also known as the Fisher discriminant [63]) of the between-class inertia B_i to the "average radius" of the scatter of all classes R_i defined as

$$r_i = \frac{B_i}{R_i} = \frac{\sum\limits_{k=1}^{K} \frac{N_k}{N} \|\mathbf{m}_{i,k} - \mathbf{m}_i\|}{\sum\limits_{k=1}^{K} \left(\frac{1}{N_k} \sum\limits_{n_k=1}^{N_k} \|\mathbf{x}_{i,n_k} - \mathbf{m}_{i,k}\| \right)}. \tag{1}$$

The principle is quite intuitive as we would like to select features that enable good separation between classes with respect to the within-class spreads. Thus, the selected additional feature corresponds to the highest ratio r_i.

Using such a criterion may result in redundant feature subsets, wherein the same signal properties are embedded in a number of features still entailing high r_i-values. Then, as described in [21], the algorithm has been modified to take into account the nonredundancy constraint by introducing an orthogonalization step at each feature selection iteration. In summary, at each iteration:

- the ratio r_i is maximized yielding a new feature subset $\mathbf{X}_i$;
- the feature space spanned by all observations is made orthogonal to $\mathbf{X}_i$.

The algorithm stops when the ratio r_d measured at iteration d gets much smaller than r_1, i.e., when $r_d/r_1 < \epsilon$ for a chosen ϵ, which means that the gain brought by the last selected feature has become nonsignificant. This provides a convenient means for implicitly selecting the number of useful features when the size of the feature subset to be selected is not a constraint.

B. Feature Selection With Genetic Algorithms (GAFS)

In this approach, the feature space is searched randomly under the guidance of a fitness function. The randomization of the search enables the algorithm to look for the features to be selected in the neighborhood of the optimal solution. Genetic algorithms belong to the family of evolutionary strategies (ES) highly inspired by natural processes [64], [65]. From an initial population of randomly generated chromosomes (each chromosome representing a candidate subset of features), a GA simulates an evolution process (which is actually a search) so that after a number of generations or iterations, the resulting more evolved chromosomes correspond to near optimal subsets of features. Evolution is represented by basic genetic operators which are fitness evaluation, selection and recombination. At each iteration, the algorithm selects the best two parent chromosomes with respect to the chosen fitness criterion for recombination. New chromosomes are thus created and integrated to the initial population. This process is repeated until some convergence condition is met. The different aspects of the algorithm we use are further explained in the following.

1) Encoding and Initialization: Chromosomes consist of binary digit strings (gene sequences) where each bit codes for the selection of a particular feature (1 for feature selected and 0 for feature not selected). The length of the chromosome is thus the total number of initial features D and each gene codes for a specific feature. At the initialization stage, chromosomes are generated randomly. Alternatively, the number of selected features can be controlled in the random generation process [62].

2) Fitness Evaluation: This is a critical operation in GAFS, since the relevancy of features being selected is measured at this stage. It is important to use fitness functions that best translate the potential classification performance resulting from the selected features. Ideally, one would use the recognition accuracies found with classification based on the considered chromosomes, but this would be computationally too expensive. The idea developed below is thus to consider more fit the feature subsets that result in the most separable class probability densities. These densities will be assumed to be Gaussian in our case.

For instance, in a 2-class situation, it is proposed to use for a chromosome C and corresponding feature subset $\mathbf{X}_C = \mathbf{X}_1^C \cup \mathbf{X}_2^C$, the fitness function F defined by

$$F(C) = J(\mathbf{X}_C) = \frac{\|\boldsymbol{\mu}_1^C - \boldsymbol{\mu}_2^C\|_2}{\sqrt{\frac{|\boldsymbol{\Sigma}_1^C| + |\boldsymbol{\Sigma}_2^C|}{2}}} \tag{2}$$

where $(\boldsymbol{\mu}_i^C)_{i=1,2}$ and $(|\boldsymbol{\Sigma}_i^C|)_{i=1,2}$ are, respectively, the mean vectors and the determinants of the diagonal covariance matrices of the multivariate Gaussian distributions that we fit to the data $\mathbf{X}_1^C$ and $\mathbf{X}_2^C$. The idea is thus to consider more fit the feature subsets that result in the most separable class probability densities which are assumed to be Gaussian.

The selection of chromosomes is then performed thanks to this fitness measure, yet it is made using probabilistic considerations. The algorithm selects the chromosomes that are *probably the most fit*. The concept is again inspired by natural processes where not necessarily the most evolved species survive into next generations, some merely have the chance to persist.

Thus, the actual selection is made by the so-called rank-based roulette-wheel rule enabling the more fit chromosomes to be more probably selected [65].

Note that we do not constrain the final subset of features to have a predetermined size. However, in order to avoid too large feature-set solutions, the fitness is penalized such that the new function F' is given by

$$F'(C) = F(C) - P(C)$$

where $P(C)$ is zero if the size of $\mathbf{X}_C$ is still smaller than a maximum chosen number and else linearly increasing with the extra number of features.

3) Crossover and Mutation: Crossover allows information exchange between two potentially fit chromosomes to give rise to a new one (an offspring) which is a hybrid version of the parents. This is how new candidate features are explored in the search space. Another genetic operator, mutation, is used to recover efficient features that could have been lost during the search. Mutation is performed with low probability as in natural processes.

C. Class Pairwise Feature Selection

Our main contribution to feature selection resides in that we perform it class pairwise. The idea is to fetch the subsets of features which are the most effective in discriminating between all possible pairs of classes. Subsequent classification is then to be performed in a one versus one scheme using as many 2-class classifiers as instrument pairs based on different feature subsets.

Not only is the approach more efficient in terms of recognition success, but also it is very convenient from an analysis point of view. In fact, it makes the optimization of classification performance more straightforward in the sense that it helps finding remedies to instrument confusions (see Section V). For example, if bad recognition accuracies are found for a given instrument i because of frequent confusions with instrument j, it is reasonable to consider optimizing only the $\{i,j\}$ classifier. In addition, better understanding of instrument timbral differences is made possible in the form of interpretations such as "*Instrument i has characteristics A and B quite different from instrument j*," where "*characteristics A and B*" are deduced from the subset of features selected for the pair $\{i,j\}$.

The pairwise solution remains practicable even when a higher number of instruments are considered since hierarchical classification, wherein instruments are grouped into families, is commonly used with success in this case [14], [19], [21]. The number of combinations to be considered at a time is then reduced to classes at the same level of taxonomy, which rarely exceed four classes.

Hereafter, we will denote classic K-class feature selection ($K > 2$) by 1-IRMFSP and use the notation C_2^K-IRMFSP and C_2^K-GAFS for pairwise feature selection. Note that in our study, genetic algorithms are only used in the class pairwise approach.

IV. Theoretical Background on Classification

A. GMMs

The GMM has been widely used in the speech/speaker community since its introduction by Reynolds for text-independent speaker identification [66]. It was also successful for musical instrument recognition [19], [30]. We give here a concise overview of the model since it is well known in the literature. In such a model, the distribution of the P-dimensional feature vectors is described by a Gaussian mixture density. For a given feature vector $\mathbf{x}$, the mixture density for the class Ω_k is defined as

$$p(\mathbf{x}|\Omega_k) = \sum_{m=1}^{M} w_{m,k} b_{m,k}(\mathbf{x}) \tag{3}$$

where the weighting factors $w_{m,k}$ are positive scalars satisfying $\sum_{m=1}^{M} w_{m,k} = 1$. The probability density is then a weighted linear combination of M Gaussian component densities $b_{m,k}(\mathbf{x})$ with mean vector $\boldsymbol{\mu}_{m,k}$ and covariance matrix $\boldsymbol{\Sigma}_{m,k}$ given by

$$b_{m,k}(\mathbf{x}) = \frac{1}{(2\pi)^{P/2}|\Sigma_{m,k}|^{1/2}} \times e^{\left(-(1/2)(\mathbf{x}-\boldsymbol{\mu}_{m,k})'(\Sigma_{m,k})^{-1}(\mathbf{x}-\boldsymbol{\mu}_{m,k})\right)}. \tag{4}$$

The parameters of the model for the class k, denoted by $\lambda_k = \{w_{m,k}, \boldsymbol{\mu}_{m,k}, \Sigma_{m,k}\}_{m=1,\ldots,M}$, are estimated using the well-known expectation-maximization (EM) algorithm [67]. Classification is usually made using the maximum *a posteriori* probability (MAP) decision rule which in virtue of Bayes rule, can be written as

$$\hat{\Omega} = \arg \max_{1 \le k \le K} \sum_{t=1}^{L} \log p(\mathbf{x}_t|\Omega_k) \tag{5}$$

where K is the number of possible classes, $p(\mathbf{x}_t|\Omega_k)$ is given in (3), $\mathbf{x}_t$ is the test feature vector observed at time t, and L is the total number of observations considered.

B. Classification by Pairwise Coupling

When addressing a K-class classification problem through multiple 2-class classifications, one is confronted with the problem of coupling the pairwise decisions at the stage of test. This issue was addressed by Hastie and Tibshirani [68] who formalized a method to perform optimal coupling.

From the set of probabilities $r_{ij} = \mathrm{Prob}(\Omega_i|\Omega_i \text{ or } \Omega_j)$ estimated for each pair $\{\Omega_i, \Omega_j\}_{1 \le i < j \le K}$ at a given observation $\mathbf{x}_t$, an estimate of the probabilities $\mathbf{p}(\mathbf{x}_t) = (p_1(\mathbf{x}_t), p_2(\mathbf{x}_t), \ldots, p_K(\mathbf{x}_t))$ is deduced assuming for r_{ij} the model

$$\mu_{ij} = \frac{p_i}{p_i + p_j} \tag{6}$$

where $p_i = \mathrm{Prob}(\Omega_i)$. The proposed algorithm finds $\mathbf{p}(\mathbf{x}_t)$ that minimizes the average weighted Kullback–Leibler distance $l(\mathbf{p})$ between r_{ij} and μ_{ij}, i.e.

$$l(\mathbf{p}) = \sum_{i<j} n_{ij} \left[r_{ij} \log\left(\frac{r_{ij}}{\mu_{ij}}\right) + (1 - r_{ij}) \log\left(\frac{1 - r_{ij}}{1 - \mu_{ij}}\right) \right] \tag{7}$$

with n_{ij} the number of training examples used to train the pair $\{\Omega_i, \Omega_j\}$ classifier. This is done by means of a gradient approach. Classification can then be made using the usual MAP decision rule [63].

When considering GMM for classification with a pairwise strategy, we use the Hastie–Tibshirani approach to couple the decisions obtained with every pair of GMM as follows. For a given test observation $\mathbf{x}_t$, and a given class pair $\{\Omega_i, \Omega_j\}$, we compute the likelihood of each class $p(\Omega_i|\mathbf{x}_t)$ and $p(\Omega_j|\mathbf{x}_t)$, and compute $\hat{r}_{ij} = p(\Omega_i|\mathbf{x}_t)/p(\Omega_i|\mathbf{x}_t) + p(\Omega_j|\mathbf{x}_t)$ and $\hat{r}_{ji} = p(\Omega_j|\mathbf{x}_t)/p(\Omega_i|\mathbf{x}_t) + p(\Omega_j|\mathbf{x}_t)$. The previous method is then used to estimate $\mathbf{p}(\mathbf{x}_t)$ assuming the model (6) for $\hat{r}_{ij}$.

C. Support Vector Machines (SVMs)

SVMs have been used successfully for various classification tasks, including speaker identification, text categorization, face recognition, etc., but also recently in musical instrument recognition [32], [33], [46]. SVMs are powerful classifiers arising from structural risk minimization theory [69] with very interesting generalization properties [70]. Another advantage of these classifiers is that they are *discriminative* by contrast to *generative* approaches (such as GMM) assuming a particular form for the data probability density which is often not consistent.

Considering two classes, SVMs try to find the hyperplane that separates the features related to each class with the maximum margin. Formally, the algorithm searches for the hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ that separates the training samples $\mathbf{x}_1, \ldots, \mathbf{x}_p$ which are assigned labels $y_1, \ldots, y_p$ ($y_i \in \{-1, 1\}$) so that

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0, \forall i \tag{8}$$

under the constraint that the distance $2/||\mathbf{w}||$ between the hyperplane and the closest sample is maximal. Vectors for which the equality in (8) holds are called support vectors.

In order to allow the algorithm to find nonlinear decision surfaces, the concept of kernel functions was introduced. Then, SVMs map the P-dimensional input feature space into a higher dimension space where the two classes become linearly separable, using a Kernel function $K(\mathbf{x_i}, \mathbf{x_j})$ such that

$$K(\mathbf{x_i}, \mathbf{x_j}) = \Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j})$$

where $\Phi : \mathrm{R}^P \longmapsto \mathrm{H}$ is a map to the high dimension space H. A great advantage of the approach resides in that one does not need to know Φ explicitly, since one only needs to know how to compute $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$; all computations can be made using the expression of $K(\mathbf{x_i}, \mathbf{x_j})$ and the problem is still solved in the low dimensional space. Interested readers are referred to [70] for further details.

SVMs are by essence 2-class classifiers. Nonetheless, they can be used to perform K-class classification using either the one versus one or one versus all strategies. In this paper, a one versus one strategy (or class pairwise strategy) is adopted and classification is then performed using a "majority vote" rule applied over all possible pairs.

V. Experimental Study

A major difficulty in the evaluation of automatic musical instrument recognition, and especially in the case where solo phrases are considered, is the lack of publicly available sound databases. As a consequence, the comparison between different proposed technologies is not straightforward. As a matter of fact, each study uses specific experimental conditions and evaluation protocols. In particular, it is important to avoid direct comparison with work on isolated notes which represents a significantly different problem.

In our work, in order to assess the generalization capability of the recognition system, a great deal of effort has been dedicated to obtain enough variation in sound material with regard to recording conditions, performers, and instrument instances.

This section presents a number of experiments to illustrate the adequacy of the feature selection (IRMFSP versus genetic algorithms), of the classification approach (GMM versus SVM) and classification strategy (K-class versus pairwise comparison) to obtain a robust musical instrument recognition system. In order to monitor the performance of our algorithm, a reference (or baseline) system has been built (see Section V-B).

A. Experimental Parameters

1) Sound Database for Solo Phrase Recognition: Ten instruments are considered, namely, alto sax, bassoon, Bb clarinet, flute, oboe, trumpet, french horn, violin, cello and piano. This choice is made so that all instrument families are represented. Moreover, potentially similar instruments (within the same family) are used so as to avoid simplification of the problem as it is much easier to discriminate the harp from the alto sax than discriminate the Bb clarinet from the alto sax, for example.

Sound samples were excerpted from compact disc (CD) recordings mainly obtained from personal collections. The content consisted of classical music and jazz from both studio and live performance, or educative material for music teaching. Additionally, alto sax, Bb clarinet and trumpet solo phrases performed by three amateur players were recorded at the Télécom Paris studio. The selection of recording excerpts used in the training set was randomly made under the constraint that at least 15 min of data could be assembled. Whenever this was not possible, at least 2 min of data were kept for testing (in the worst case) and the rest was used for training in order to provide tight confidence ranges on the estimation of recognition accuracies. Ideally, never would the same CD-recording provide excerpts for both training and test sets, but in some cases, it was not possible to do so without lacking of material either for training or testing. However, it was made sure that samples used for testing were never extracted from tracks whose any part was included in the training set. Table I sums up the properties of the data used in our experiments. The diversity of the sound database properties used in studies on instrument recognition on solo phrases (including ours) is illustrated in Table II which highlights the difficulty to directly compare their respective performances.

2) Signal Processing: Previous work on instrument recognition has shown that a 32-kHz sampling frequency is not penalizing for classification performance [30], which led us to down-sample the input signal to this frequency in order to reduce the computational load. Additionally, the signal was centered with respect to its temporal mean, and its amplitude was normalized with respect to its maximum value. The analysis was performed over sliding overlapping windows. The frame length was 32 ms and the hop size was 16 ms for the extraction

TABLE I
SOUND DATABASE – *SOURCES*, *TRACKS*, AND *FRAME NBR* ARE, RESPECTIVELY, THE TOTAL NUMBER OF DISTINCT SOURCES, THE TOTAL NUMBER OF TRACKS FROM CDs, AND THE NUMBER OF 32-ms TEST FRAMES USED FOR TEST; *TOTAL TRAIN* AND *TOTAL TEST* ARE THE TOTAL DURATIONS OF, RESPECTIVELY, TRAIN AND TEST MATERIAL IN SECONDS

	Total train (s)	Sources	Tracks	Frame nbr	Total test (s)
Alto Sax	523	10	19	19800	310
Bassoon	176	5	9	8280	130
Bb Clarinet	756	10	26	31140	488
Flute	606	8	24	44190	692
Oboe	1074	8	24	71310	1117
French Horn	261	5	13	7050	110
Trumpet	1158	9	73	66960	1049
Cello	1101	7	20	65490	1026
Violin	1325	11	31	59790	937
Piano	1203	8	15	45870	719

TABLE II
SOUND DATABASE – *CLASSES* IS THE NUMBER OF INSTRUMENT CLASSES STUDIED (WHEN AT LEAST TWO INSTANCES WERE AVAILABLE) *SOURCES* IS THE NUMBER OF DISTINCT SOURCES USED; *TRAIN* AND *TEST* ARE, RESPECTIVELY, THE TOTAL LENGTH OF THE TRAINING DATA, AND TOTAL LENGTH OF TEST DATA, IN SECONDS; MINIMUM AND MAXIMUM DURATIONS ARE GIVEN

	Classes	*Sources*	*Train* (s)	*Test* (s)
Brown [30]	4	-	54 - 330	60 - 240
Martin [14]	11	2 - 8	12 - 2130	54 - 2130
Marques [46]	8	2 - 2	205 - 205	20 - 20
Miravet [31]	6	3 - 9	1818 - 2044	945 - 1136
This work	10	**5 - 11**	176 - 1325	110 - 1117

TABLE III
FEATURE SUBSETS AND THEIR CODES (COLUMN ONE); FEATURE SUBSET SIZES (COLUMN TWO); FEATURES SELECTED USING 1-IRMFSP (COLUMN THREE)

Feature subset	Size	Selected
AC=[AC1,...,AC49]	49	-
ZCR	1	-
MFCC=[C1,..,C10]+δ+δ^2	30	C1,...,C4
Sx=[Sc,Sw,Sa,Sf]+δ+δ^2	12	Sc,Sw,Sa,Sf
ASF=[A1,...,A23]	23	A22,A23
Si=[S1,...,S21]	21	-
Fc	1	-
OBSI=[O1,...,O8]	8	O4,...,O8
OBSIR=[OR1,...,OR7]	7	OR4,...,OR7
AM=[AM1,...,AM8]	8	-

TABLE IV
BASELINE SYSTEM: 10-CLASS GMM CLASSIFICATION WITH 1-IRMFSP (COLUMN TWO); ONE VERSUS ONE GMM CLASSIFICATION WITH 1-IRMFSP (COLUMN THREE) AND C_2^{10}-IRMFSP (COLUMN FOUR)

	1-IRMFSP 10-class	1-IRMFSP 1 vs 1	C_2^{10}-IRMFSP 1 vs 1
Alto Sax	61	62	73
Bassoon	68	68	60
Bb Clarinet	71	73	79
Flute	80	80	88
Oboe	75	75	78
French Horn	55	55	76
Trumpet	82	83	85
Cello	88	88	94
Violin	89	88	90
Piano	82	82	98
Average	75	75	82

of all features except tremolo and roughness. Longer analysis length (960 and 480-ms hopsize) was used for the latter so as to measure the AM features properly. All spectra were computed with a fast Fourier transform (FFT) after Hamming windowing. Frames consisting of silence signal were detected thanks to a heuristic approach based on power thresholding then discarded from both train and test data sets. The frequency ratio for the constant-Q transform was 1.26. A total of 160 feature coefficients were considered including elements from all feature subsets described earlier.

All features were rescaled in order to homogenize the highly varying dynamics of the different feature subsets in such a way that all coefficients were confined in the range [0,1]. This is done by normalizing the features with respect to scale factors deduced from their "ceiled" maximum values (estimated during training). Such a preprocessing has proven to be successful for better classification [34].

B. Baseline System

The baseline system follows a classic K-class GMM approach where the model orders M_k for each class k vary in the set {8,16,32,64,128,256} and are selected using a Bayesian information criterion (BIC) [71]. For this reference system, 1-IRMFSP was used for feature selection, and a MAP criterion used for decision. Scoring was performed as follows: for each test signal, a decision regarding the instrument classification was taken every $T = 0.47$ s ($L = 30$ overlapping frames of 32-ms duration). The recognition success rate is then, for each instrument, the percentage of successful decisions over the total number of T-second test segments.

The results of this baseline system obtained on our database are given in column two of Table IV. The average accuracy is 75%. Although acceptable results are obtained for some instruments as the violin for example (89%), the recognition of others remains unsatisfactory (as for the french horn successfully classified only 55% of the time). We will show that the average accuracy can be improved with our approach.

C. Experiment 1, Feature Selection

1) K*-Class Feature Selection:* An overview of the different feature subsets used in our experiments is presented in Table III together with the 19 features selected through the 1-IRMFSP approach (column three) using a convergence condition determined by $\epsilon = 10^{-5}$. The efficiency of the OBSI/OBSIR attributes is confirmed since they are largely represented in the subset of selected features. Features describing the spectral shape (Sc, Sw, Sa, Sf) as well as ASF coefficients were found very useful. Only the first four MFCCs were selected.

2) K*-Class Feature Selection and Pairwise Classification:* Column three of Table IV provides the recognition accuracies obtained with 1-IRMFSP and a one versus one GMM classification (as described in Section IV-B). It can be noticed that the pairwise classification does not bring any significant improvement compared to the reference system.

3) Pairwise Feature Selection and Pairwise Classification: Recognition accuracies obtained with one versus one GMM classification based on C_2^{10}-IRMFSP are given in column four of Table IV. Substantial improvement in recognition accuracy (up to $+22\%$ for the french horn) is achieved with

TABLE V
FEATURES SELECTED BY THE C_2^{10}-IRMFSP ALGORITHM FOR A FEW EXAMPLES. "FR" STANDS FOR FREQUENCY AND "ST" FOR STRENGTH

Bb Clarinet/Alto Sax	Bb Clarinet/Bassoon	Bb Clarinet/Flute	Bb Clarinet/French Horn
C1,..,C3,C6,..,C8,δC0 Sw,Sa,Sf, A5, A9, A10 A12, A15, A19, A20, A21, A22, A23, AM fr x st 4-8Hz AM fr 10-40Hz, AM st 10-40Hz, Fc, OR2, OR5, OR6, OR7, S8, S14	C1,..,C4 Sc, Sw, Sa A21, A22, A23 OR5, OR6, OR7 S12, S18	C1, C2, C3, C6, δ^2C0, Sc, δ^2Sc, Sa, Sf, A5, A9, A10, A18, A20, A22, A23, AM fr 10-40Hz, AC5, AC10, AC23, AC42, Fc, ZCR, O1, O2, O3, O4, O5, O6, O7, O8, OR1, OR2, OR3, OR4, OR5, OR6, OR7, S7, S8, S15, S16, S18, S19	C1, C2, C3, C4, C5, C6, Sc, Sw, Sa, Sf, A2, A3, A5, A6, A9, A10, A14, A18, A20, A23, AM fr 4-8Hz, AM st 4-8Hz, AM heur st 4-8Hz, AM st 10-40Hz, Fc, ZCR, OR5, OR6, S9, S13, S14, S15, S16, S20.

Bb Clarinet/Trumpet	Bb Clarinet/Cello	Bb Clarinet/Violin	Bb Clarinet/Piano	Bb Clarinet/Oboe
C2, C3, Sw, Sa, Sf, AC8, O1, O5, O6, O7, OR5, OR7, S15, S16, S19,	C1, C2, C3, Sw, Sa, Sf, A22, AM fr 4-8Hz, AC1, O5, OR1, S19	C1, C2, C3, Sw, Sa, Sf, A20, A22, A23, Fc, O4, O5	C1, C2, C3, C4, Sw, Sa, A13, A18, A20, A22, A23, AM frequency 4-8Hz, AC1, Fc, O2, O6, O7, O8, OR6, OR7.	C2, C3, C4, C5, C7, Sc, Sw, Sa, A22, AC1, AC8, AC18, O2, O4, O6, O7, O8, OR5, OR7, S11, S14

C_2^{10}-IRMFSP for all instruments except the bassoon. The average improvement is seven percentage points.

Note that, for C_2^{10}-IRMFSP, a different model is computed for the same instrument class C_i with respect to the instrument class C_j it is confronted with, since a specific subset of features is selected for the pair (C_i, C_j). The model order M_{ij} of each GMM is also assessed using a BIC approach with $M_{ij} \in \{8,16,32,64,128,256\}$.

Pairwise IRMFSP was performed (with the same convergence criterion $\epsilon = 10^{-5}$). On average, the same number of features (19) is selected. While the same feature subsets (OBSI/OBSIR, Sc, Sw, Sa, Sf, ASF) remain the most efficient, more features are selected by the algorithm for specific pair combinations where more attributes are necessary for better discrimination. Spectral "irregularity" coefficients (Si) were considered particularly useful for combinations involving the Bb clarinet versus another wind instrument and otherwise rarely selected. AM features were particularly consistent when dealing with wind instruments, especially with the Bb clarinet and the french horn. A maximum of four autocorrelation coefficients (among 49) were selected for the pair Bb clarinet/flute. Zero crossing rate was selected 18 times (out of 45) and frequency cutoff 21 times. As for delta-cepstrum attributes, only energy temporal variation (δC0) and energy acceleration (δ^2C0) were found efficient for only a few combinations. On the contrary, in other cases, a number of features are found not useful for given instrument pairs; hence, they are not selected. This results in sizes of selected feature subsets ranging from nine (for the piano/violin pair, for which only the three first MFCC, the spectral moments and the fifth and eighth OBSI coefficients were selected) to 44 (for Bb clarinet versus flute). Examples of class pairwise feature selection results are presented in Table V. All selected feature subsets were posted on the web [72] for interested readers to look into it in depth.

4) Genetic Algorithms for Feature Selection: A tentative to improve feature selection was made using genetic algorithms also performed in a pairwise fashion (denoted by C_2^{10}-GAFS). We use the fitness measure described in Section III-B. Two variants are tested: a classic one with totally random initialization and an alternative approach with assisted initialization, where we introduce an evolved chromosome in the initial population, among the randomly generated other initial chromosomes, in order to obtain a set of features more fit than the IRMFSP one. This is achieved by introducing at the initialization stage a chromosome constructed with genes obtained with the C_2^{10}-IRMFSP algorithm findings (with $\epsilon = 10^{-5}$).

The GAFS algorithm often introduced autocorrelation coefficients (AC) in the subset of the most relevant features. These were hardly selected by IRMFSP. The average number of selected features is 33.

To test the performance of feature selection algorithms, basic linear SVM classification is used. Recognition accuracies thus found are presented in Table VI. Note that these results are to be compared intrinsically rather than with Table IV. IRMFSP is tested with two stop criteria, $\epsilon = 10^{-5}$ [column two, denoted by IRMFSP (10^{-5})] resulting in an average of 19 selected features and $\epsilon = 10^{-6}$ [column three, denoted by IRMFSP 10^{-6})] for 38 selected features on average. Results obtained with classic GAFS and GAFS with assisted initialization are given, respectively, in columns four and five. As expected, IRMFSP (10^{-6}) provides the best overall performance since more features are selected on average. The average improvement in recognition accuracies is 4% compared to IRMFSP (10^{-5}). Although the average recognition rate is 73% with features selected using GAFS with random initialization, this algorithm remains less efficient than IRMFSP (10^{-6}) except for the recognition of the oboe, the trumpet and the violin. When testing GAFS with assisted initialization, some improvement is often observed compared to IRMFSP (10^{-5}) yet more features are selected and this approach performs better than IRMFSP(10^{-6}) only for alto sax. It is believed that the used fitness measure was not always optimal because it is based on the assumption that the data has Gaussian distribution (see Section III-B2). As a result, the selected set of features, although fit with respect to the chosen fitness function, do not satisfy the properties we are requiring. This confirms the importance of a judicious choice of the fitness function to be used in GAFS. A promising candidate, that is being studied, is the $\xi\alpha$ estimate of the SVM classifier success [73].

5) Optimization of Feature Selection by Fusion: This situation allows us to show the flexibility of the pairwise classification approach. A major advantage is that we can still exploit only the improved feature subsets in order to optimize a classification system performing better than the one using IRMFSP (10^{-6}), by altering only a few classifiers among all the pairs.

TABLE VI
CLASSIFICATION PERFORMANCE WITH C_2^{10}-IRMFSP (COLUMNS TWO AND THREE), C_2^{10}-GAFS (COLUMNS FOUR AND FIVE)

	IRMFSP (10^{-5})	IRMFSP (10^{-6})	GAFS	GAFS (init+)
Alto Sax	43	50	40	54
Bassoon	51	59	35	52
Bb Clarinet	81	86	73	82
Flute	76	84	63	80
Oboe	76	78	78	74
French Horn	71	75	69	66
Trumpet	84	86	88	86
Cello	94	95	92	88
Violin	91	94	94	93
Piano	99	99	98	97
Average	77	81	73	77

TABLE VII
PARTIAL CONFUSION MATRICES FOR CLASSIFICATIONS, FROM LEFT TO RIGHT BASED ON C_2^{10}-GAFS – C_2^{10}-IRMFSP AND (OPTIMIZED FEATURE SETS). READ ROW CONFUSED WITH COLUMN

	Alto Sax	Bb Clarinet	Violin	Trumpet
Alto Sax	*54 - 50 (56)*	6 - 6 (6)	31 - 35 (29)	4 - 3 (4)
Bb Clarinet	0 - 0 (0)	*82 - 86 (87)*	1 - 2 (2)	4 - 3 (3)
Violin	3 - 3 (2)	1 - 2 (2)	*94 - 94 (94)*	2 - 1 (1)
Trumpet	0 - 3 (1)	0 - 1 (1)	3 - 3 (4)	*88 - 86 (87)*

The following example is illustrating the procedure. Looking at the confusions made by the classification based on C_2^{10}-GAFS and C_2^{10}-IRMFSP (10^{-6}) (given in Table VII), one can work out that the alto sax was confused with the violin 35% of the time with IRMFSP, and only in 31% of the cases using GAFS.

Thus, we replace the feature subset found by IRMFSP by the one found with GAFS for the discrimination between the pair (alto sax, violin) which results in smaller confusion between these two instruments compared to the results with IRMFSP (alto sax is now confused with violin 29% of the time). The same process is repeated for all situations where GAFS provides better discrimination between a pair of instruments, yielding a hybrid set of features consisting of pairwise chosen subsets compiled from the best of C_2^{10}-GAFS and C_2^{10}-IRMFSP. Preliminary results, found using the same test set, show that some improvement of the recognition accuracy (compared to the original found by C_2^{10}-IRMFSP) can thus be achieved.[2] The optimization is not always successful since all confusions should be optimized jointly. In fact, a given feature subset may result in instrument i being less confused with instrument j and at the same time j being more confused with i (see the confusions for the pair (alto sax, trumpet) for example). Nonetheless, substantial improvement is achieved for individual instrument classes using an optimization that is not practicable in a ten-class classification scheme wherein a unique set of features is used that cannot be altered without changing all recognition accuracies.

D. *Experiment 2, SVM Kernels*

For all the following experiments, we keep unchanged the features selected pairwise in the previous experiments (those compiled from C_2^{10}-GAFS and C_2^{10}-IRMFSP (10^{-6}) in Section V-C5) to study aspects related to classification.

TABLE VIII
CLASSIFICATION RESULTS WITH SVM USING LINEAR, POLYNOMIAL ($d = 2, \ldots, 5$) AND RBF KERNELS WITH OPTIMIZED FEATURE SUBSETS. BEST SCORES ARE GIVEN IN BOLD

	Linear	Poly(d=2)	Poly(d=3)	Poly(d=4)	Poly(d=5)	RBF
Alto Sax	56	61	64	64	64	**70**
Bassoon	59	66	66	**67**	66	66
Bb Clarinet	87	91	93	94	94	**96**
Flute	84	88	89	90	90	**92**
Oboe	79	81	82	82	82	**83**
French Horn	74	78	78	78	78	**81**
Trumpet	87	88	89	89	89	**90**
Cello	95	**96**	**96**	**96**	**96**	**96**
Violin	94	**96**	**96**	**96**	**96**	**96**
Piano	99	**100**	**100**	**100**	**100**	**100**
Average	81	84	85	85	85	**87**

We examine here the efficiency of SVM classification for musical instrument recognition on solo phrases using different kernels. Three types of kernel are examined, linear (or no kernel), polynomial, and radial basis function (RBF). The used polynomial kernel has the form

$$K(\mathbf{x}, \mathbf{y}) = (s\, \mathbf{x} \cdot \mathbf{y} + c)^d.$$

As for the RBF kernel, it is given by

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{y}\|^2\right).$$

The recognition accuracies obtained with the different kernels are given in Table VIII. The RBF kernel is the most successful with an average accuracy of 87%. When using the polynomial kernel, increasing the degree from 2 to 4 results in increased performance. A degree greater than 4 is not efficient since the performance remains unchanged for increased computational load. The fourth-degree polynomial kernel is the most interesting polynomial kernel as it results in the best individual and average accuracies and performs better than the RBF kernel for the recognition of the bassoon. It is worth to note that the piano is very easily discriminated from other instruments since its recognition accuracy is already 99% without any kernel. Finally, note that GMM were more successful for the recognition of the alto sax (73% with GMM). The previous thus suggests combining the different classifiers [74] for better overall performance.

E. *Experiment 3, Changing the Decision Length*

The last experiment is concerned with the influence of the decision length on the recognition accuracy. So far, $L = 30$ successive overlapping 32-ms frames have been considered in classifying a given test signal i.e., the decision length has been 0.47 s. Table IX presents the recognition accuracies obtained using longer decision lengths.

We considered the cases $L = 60$ ($\approx$ 1 s) and $L = 320$ ($\approx$ 5 s). High accuracies are found. The average is 88% with 1-s segments ($L = 60$) and 93% with 5-s segments ($L = 320$). The recognition of the piano is always successful from 0.5-s decision lengths on and so it is for the Bb clarinet with 5-s decisions.

[2] these results are considered as preliminary since we unfortunately lack a development set to be used to perform the optimization of the features selected. This led us to exploit, in the optimization, the confusions found over the test set, to illustrate the proposed procedure.

TABLE IX
CLASSIFICATION PERFORMANCE FOR DIFFERENT DECISION LENGTHS USING THE OPTIMIZED FEATURE SUBSETS AND SVM WITH A RBF KERNEL

	$L\approx$0.5s (30)	$L\approx$1s (60)	$L\approx$5s (320)
Alto Sax	70	73	82
Bassoon	66	67	82
Bb Clarinet	96	98	100
Flute	92	92	95
Oboe	83	84	83
French Horn	81	84	94
Trumpet	90	91	93
Cello	96	96	98
Violin	96	96	99
Piano	100	100	100
Average	87	88	93

VI. CONCLUSION

Machine recognition of musical instruments on solo performance has been addressed. A number of potentially useful signal processing features have been studied. New features were proposed, namely octave band signal intensities and octave band signal intensity ratios that prove highly efficient for the recognition task. inertia ratio maximization using feature space projection and genetic algorithms have been considered for feature selection.

Moreover, we have shown that it is very advantageous to perform feature selection class pairwise, looking for the subsets of features that enable the best discrimination between any possible pair of instrument classes. It entails much better recognition accuracies and allows us to optimize simple 2-class schemes for better overall performance. Furthermore, it is an interesting starting point for studying timbral differences between instruments. In fact, it guides one to natural formulations of the relations existing among them by establishing simple binary comparisons. Nevertheless, some higher level characterization of the selected low-level features is needed to gain better understanding of these relations.

Two types of classifiers were studied, GMM and SVM, that were exploited in a one versus one scheme. SVM with a RBF kernel gave the best results (on average 12% improvement was achieved compared to our baseline system). Further improvement of the recognition accuracies was obtained using a larger number of observations for decisions, which resulted in high recognition performance (93%).

Future work will consider alternative feature selection techniques better adapted to SVM classification. Furthermore, hierarchical classification wherein instruments are grouped into families will be envisaged. The recognition of typical instrumental ensembles (solos, duets, trios, etc.) will be introduced at a high level of taxonomy. As for classification, probabilistic outputs for SVM will be considered together with a time dynamic approach.

REFERENCES

[1] *Information Technology – Multimedia Content Description Interface – Part 4: Audio, Int. Standard*, ISO/IEC FDIS 15938–4:2001(E), Jun. 2001.

[2] D. Godsmark and G. J. Brown, "A blackboard architecture for computational auditory scene analysis," *Speech Commun.*, vol. 27, pp. 351–366, Oct. 1998.

[3] K. Kashino and H. Mursae, "A sound source identification system for ensemble music based on template adaptation and music stream exrtaction," *Speech Commun.*, vol. 27, pp. 337–349, Sep. 1998.

[4] T. Kinoshita, S. Sakai, and H. Tanaka, "Musical sound source identification based on frequency component adaptation," in *Proc. IJCAI Workshop on Computational Auditory Scene Analysis (IJCAI-CASA)*, Stockholm, Sweden, Aug. 1999, pp. 18–24.

[5] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China, Apr. 2003, pp. 553–556.

[6] ——, "Instrument recognition in accompanied sonatas and concertos," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Montréal, QC, Canada, May 2004, pp. 217–220.

[7] E. Vincent and X. Rodet, "Instrument identification in solo and ensemble music using independent subspace analysis," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Barcelona, Spain, Oct. 2004.

[8] S. Essid, G. Richard, and B. David, "Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies," Trans. Audio, Speech, and Lang. Process., vol. 14, no. 1, pp. 68–80, Jan. 2006, to be published.

[9] M. Clark, P. Robertson, and D. A. Luce, "A preliminary experiment on the perceptual basis for musical instrument families," *J. Audio Eng. Soc.*, vol. 12, pp. 199–203, 1964.

[10] R. Plomp, "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, R. Plomp and G. Smoorenburg, Eds. Leiden, Germany: Sijthoff, 1970, pp. 197–414.

[11] K. M. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1270–1277, 1977.

[12] R. A. Kendall, "The role of acoustic signal partitions in listener categorization of musical phrases," *Music Perception*, vol. 4, pp. 185–214, 1986.

[13] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres : Common dimensions, specificities and latent subject classes," *Psychol. Res.*, vol. 58, pp. 177–192, 1995.

[14] K. D. Martin, "Sound-source recognition: A theory and computational model," Ph.D dissertation, Media Lab., Dept. Elect. Eng. Comput. Sci., Mass. Inst. Technol., Cambridge, 1999.

[15] I. Kaminskyj, "Multi-feature musical instrument sound classifier," in *Proc. Australasian Computer Music Conf.*, Jul. 2000, pp. 53–62.

[16] I. Fujinaga and K. MacMillan, "Realtime recognition of orchestral instruments," in *Proc. Int. Computer Music Conf.*, 2000, pp. 141–143.

[17] B. Kostek and A. Czyzewski, "Automatic recognition of musical instrument sounds – Further developments," in *Proc. 110th AES Convention*, Amsterdam, The Netherlands, May 2001.

[18] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," *Proc. EURASIP J. Appl. Signal Process.*, vol. 1, no. 11, pp. 5–14, 2003.

[19] A. Eronen, "Automatic musical instrument recognition," Master's thesis, Dept. Inf. Tech., Tampere Univ. Technol., Tampere, Finland, 2001.

[20] ——, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs," in *Proc. 7th Int. Symp. Signal Processing and its Applications*, Paris, France, Jul. 2003, pp. 133–136.

[21] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization," in *Proc. 115th AES Convention*, New York, Oct. 2003.

[22] A. G. Krishna and T. V. Sreenivas, "Music instrument recognition : From isolated notes to solo phrases," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Montréal, QC, Canada, May 2004, pp. 265–268.

[23] F. Opolko and J. Wapnick, *McGill University Master Samples*. Montréal, QC, Canada: McGill Univ., 1987.

[24] [Online]. Available: http://theremin.music.uiowa.edu

[25] [Online]. Available: http://www.ircam.fr

[26] M. Goto, H. Hashigushi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002.

[27] O. Gillet and G. Richard, "Automatic transcription of drum loops," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Montréal, QC, Canada, May 2004, pp. iv 263–iv 272.

[28] S. Dubnov and X. Rodet, "Timbre recognition with combined stationary and temporal features," in *Proc. Int. Computer Music Conf.*, 1998.

[29] J. C. Brown, “Computer identification of musical instruments using pattern recognition with cepstral coefficients as features,” *J. Acoust. Soc. Amer.*, vol. 105, pp. 1933–1941, Mar. 1999.

[30] J. C. Brown, O. Houix, and S. McAdams, “Feature dependence in the automatic identification of musical woodwind instruments,” *J. Acoust. Soc. Amer.*, vol. 109, pp. 1064–1072, Mar. 2000.

[31] R. Ventura–Miravet, F. Murtagh, and J. Ming, “Pattern recognition of musical instruments using hidden Markov models,” in *Proc. Stockholm Music Acoustics Conf.*, Stockholm, Sweden, Aug. 2003, pp. 667–670.

[32] S. Essid, G. Richard, and B. David, “Musical instrument recognition on solo performance,” in *Proc. Eur. Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sep. 2004, pp. 1288–1292.

[33] ——, “Efficient musical instrument recognition on solo performance music using basic features,” in *Proc. AES 25th Int. Conf.*, London, U.K., Jun. 2004.

[34] ——, “Musical instrument recognition based on class pairwise feature selection,” in *Proc. 5th Int. Conf. Music Information Retrieval (ISMIR)*, Barcelona, Spain, Oct. 2004.

[35] A. Livshin and X. Rodet, “Musical instrument identification in continuous recordings,” in *Proc. 7th Int. Conf. Digital Audio Effects (DAFX-4)*, Naples, Italy, Oct. 2004, pp. 222–227.

[36] ——, “Instrument recognition beyond separate notes – Indexing continuous recordings,” in *Proc. Int. Computer Music Conf.*, Miami, FL, Nov. 2004.

[37] A. Eronen, “Comparison of features for musical instrument recognition,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2001, pp. 19–22.

[38] G. Peeters and X. Rodet, “Automatically selecting signal descriptors for sound classification,” in *Proc. Int. Computer Music Conf.*, Goteborg, Sweden, Sep. 2002.

[39] P. Herrera, G. Peeters, and S. Dubnov, “Automatic classification of musical sounds,” *J. New Music Res.*, vol. 32, no. 1, pp. 3–21, 2003.

[40] I. Kaminskyj and A. Materka, “Automatic source identification of monophonic musical instrument sounds,” in *Proc. IEEE Int. Conf. Neural Netw.*, 1995, pp. 189–194.

[41] I. Fujinaga, “Machine recognition of timbre using steady-state tone of acoustic musical instruments,” in *Proc. Int. Computer Music Conf.*, 1998.

[42] G. Agostini, M. Longari, and E. Pollastri, “Musical instrument timbres classification with spectral features,” in *Proc. Int. Workshop Multimedia Signal Processing*, Cannes, France, Oct. 2001, pp. 97–102.

[43] B. Kostek, *Soft Computing in Acoustics, Applications of Neural Networks, Fuzzy Logic, and Rough Sets to Musical Acoustics, Studies in Fuzziness and Soft Computing*. New York: Physica-Verlag, 1999.

[44] T. Kitahara, M. Goto, and H. G. Okuno, “Musical instrument identification based on f0-dependent multivariate normal distribution,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China, Apr. 2003, pp. 409–412.

[45] J. Lee and J. Chun, “Musical instrument recognition using hidden Markov model,” in *Conf. Rec. 36th Asilomar Conf. Signals, Systems, and Computers*, Nov. 2002, pp. 196–199.

[46] J. Marques and P. J. Moreno, “A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines,” Compaq Computer Corp., Cambridge, MA, Tech. Rep. CRL 99–4, 1999.

[47] C. L. Krumhansl, “Why is musical timbre so hard to understand? in structure and perception of electroacoustic sound and music,” *Excerpta Medica*, no. 846, pp. 43–53, 1989.

[48] G. Peeters, “A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project,” IRCAM, Paris, France, Tech. Rep., 2004.

[49] F. Pachet and A. Zils, “Evolving automatically high-level music descriptors from acoustic signals,” in *Proc. 1st Int. Symp. Computer Music Modeling and Retrieval (CMMR)*, Montpellier, France, May 2003.

[50] S. Essid, P. Leveau, G. Richard, L. Daudet, and B. David, “On the usefulness of differentiated transient/steady-state processing in machine recognition of musical instruments,” in *Proc. AES 118th Convention*, Barcelona, Spain, May 2005.

[51] J. Brown, “Musical instrument identification using autocorrelation coefficients,” in *Proc. Int. Symp. Musical Acoustics*, 1998, pp. 291–295.

[52] L. R. Rabiner, *Fundamentals of Speech Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993, Prentice-Hall Signal Processing Series.

[53] M. Partridge and M. Jabri, “Robust principal component analysis,” in *Proc. IEEE Signal Processing Society Workshop*, Dec. 2000, pp. 289–298.

[54] R. Kohavi and G. John, “Wrappers for featue subset selection,” *Artif. Intell. J.*, vol. 97, no. 1–2, pp. 273–324, 1997.

[55] A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artif. Intell. J.*, vol. 97, no. 1–2, pp. 245–271, Dec. 1997.

[56] I. Guyon and A. Elisseeff, “An introduction to feature and variable selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[57] W. Sidelcki and J. Sklansky, “A note on genetic algorithms for large-scale feature selection,” *Pattern Recognit. Lett.*, vol. 10, pp. 335–347, 1989.

[58] F. J. Ferri, V. Kadirakamanathan, and J. Kittler, “Feature subset search using genetic algorithms,” in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, 1993.

[59] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, “Comparative study of techniques for large-scale feature selection,” *Pattern Recognit. Practice IV*, pp. 403–413, 1994.

[60] L. Y. Tseng and S. B. Yang, “Genetic algorithms for clustering, feature selection and classification,” in *Proc. IEEE Int. Conf. Neural Networks*, Jun. 1997, pp. 1612–1616.

[61] Z. Sun, X. Yuan, and S. J. Louis, “Genetic feature subset selection for gender classification,” in *Proc. 6th IEEE Workshop on Applications of Computer Vision*, 2002, pp. 165–170.

[62] I.-S. Oh, J.-S. Lee, and B.-R. Moon, “Local search-embedded genetic algorithms for feature selection,” in *Proc. 16th IEEE Int. Conf. Pattern Recognition*, 2002, pp. 148–151.

[63] R. Duda and P. E. Hart, *Pattern Classification and Scence Analysis*. New York: Wiley, 1973.

[64] M. Berthold and D. J. Hand, *Intelligent Data Analysis: An Introduction*. New York: Springer-Verlag, 1999.

[65] M. Srinivas and M. Patnaik, “Genetic algorithms : A survey,” in *IEEE Computer Society Press* Washington, DC, Jun. 1999, vol. 27, pp. 17–26.

[66] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[67] T. K. Moon, “The expectation-maximization algorithm,” *IEEE Signal Process. Mag.*, no. 11, pp. 47–60, Nov. 1996.

[68] T. Hastie and R. Tibshirani, ““Classification by Pairwise Coupling,” Tech. Rep.,” Stanford Univ./Univ. Toronto, 1996.

[69] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

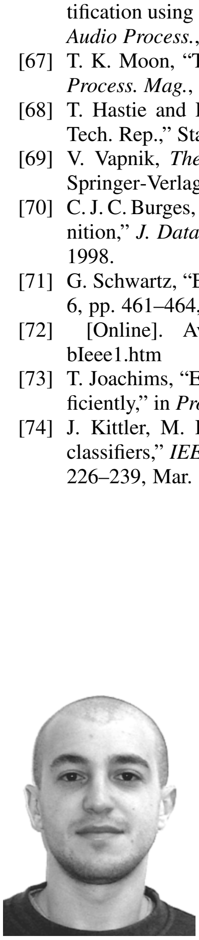
[70] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *J. Data Mining Knowledge Discovery*, vol. 2, no. 2, pp. 1–43, 1998.

[71] G. Schwartz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[72] [Online]. Available: http://www.tsi.enst.fr/essid/pub/pubIeee1/pubIeee1.htm

[73] T. Joachims, “Estimating the generalization performance of a SVM efficiently,” in *Proc. Int. Conf. Machine Learn.*, 2000, pp. 431–438.

[74] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.



Slim Essid received the electrical engineering degree from the Ecole Nationale d’Ingénieurs de Tunis, Tunis, Tunisia, in 2001 and the D.E.A (M.Sc.) degree in digital communication systems from the Ecole Nationale Supérieure des Télécommunications (ENST), the Université Pierre et Marie Curie (Paris VI), and the Ecole Supérieure de Physique et de Chimie Industrielle, Paris, France, in 2002. As part of his Master’s thesis work, he was involved in a National Telecommunication Research Network (RNRT) project to propose a low bitrate parametric audio coding system for speech and music. He is currently pursuing the Ph.D. degree at the Department of Signal and Image Processing, ENST, Université Pierre et Marie Curie with a thesis on music information retrieval.

Gaël Richard (M'02) received the state engineering degree from the Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1990, the Ph.D. degree from LIMSI-CNRS, University of Paris-XI, in 1994 in the area of speech synthesis, and the Habilitation à Diriger des Recherches degree from the University of Paris XI in September 2001.

After the completion of the Ph.D. degree, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the speech processing group of Prof. J. Flanagan, where he explored innovative approaches for speech production. From 1997 and 2001, he successively worked for Matra Nortel Communications and for Philips Consumer Communications. In particular, he was the Project Manager of several large-scale European projects in the field of multimodal verification and speech processing. In 2001, he joined the Department of Signal and Image Processing, ENST, as an Associate Professor in the field of audio and multimedia signals processing. He is coauthor of over 50 papers and inventor in a number of patents, he is also one of the experts of the European commission in the field of man/machine interfaces.

Bertrand David was born in Paris, France, on March 12, 1967. He received the M.Sc. degree from the University of Paris-Sud, in 1991 and the Agrégation, a competitive french examination for the recruitment of teachers, in the field of applied physics, from the Ecole Normale Supérieure (ENS), Cachan, France, and the Ph.D. degree from the University of Paris VI in 1999 in the field of musical acoustics and signal processing.

From 1996 to 2001, he was a Lecturer in the graduate school in electrical engineering, computer science, and communications. He is now an Associate Professor with the Department of Signal and Image Processing, Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France. His research interests include parametric methods for the analysis/synthesis of musical signals and parameter extraction for music description and musical acoustics.