MUSICAL INSTRUMENT RECOGNITION BASED ON CLASS PAIRWISE FEATURE SELECTION

Slim Essid, Gael Richard, Bertrand David

MUSICAL INSTRUMENT RECOGNITION BASED ON CLASS PAIRWISE FEATURE SELECTION

Slim ESSID, Gaël RICHARD and Bertrand DAVID
GET-ENST (Télécom Paris)
Département Traitement du Signal et des Images (TSI)

ABSTRACT

In this work, musical instrument recognition is considered on solo music from real world performance. A large sound database is used that consists of musical phrases excerpted from commercial recordings with different instrument instances, different players, and varying recording conditions.
The proposed recognition scheme exploits class pairwise feature selection based on inertia ratio maximization. Moreover, new signal processing features based on octave band energy measures are introduced that prove to be useful. Classification is performed using Gaussian Mixture Models in a one vs one fashion in association with a data rescaling procedure as pre-processing.
Experimental results show that substantial improvement in recognition success is thus achieved.

1. INTRODUCTION

Musical instrument recognition is an important aspect of music information retrieval. Such a capability may be extremely helpful in the framework of automatic musical transcription systems as well as in content-based search applications. Both the amateur music lover and the professional musician would appreciate to have a system informing them of the instruments involved in the musical piece which they are listening to.

However, processing complex mixtures of instruments of real world music remains a very difficult issue which has been barely addressed. In fact, most effort was dedicated to musical instrument recognition based on isolated-note content, and to a smaller extent, based on monophonic musical phrases [1]. In our work, music from real solo performance is considered since it is believed that this direction could give rise to immediate applications and stands as an important intermediate step towards musical recognition in the polyphonic context [2].

In marked contrast to other pattern recognition tasks (typically speech recognition), there has been no real consensus in choosing a set of signal processing features amenable to successful instrument recognition. A large number of potentially useful features can be chosen which are adapted to our task. In such a situation, feature selection techniques should be considered [3] in order to fetch the most relevant feature subset. Classically, features from all instrument classes are processed jointly, which results in an optimal set of descriptors that is subsequently used to train appropriate classifiers [4, 5, 6]. Our contribution suggests performing class pairwise feature selection in order to find the most efficient features in discriminating between a given comparison of 2 instruments. It is shown that combining this approach with a one vs one classification strategy based on Gaussian Mixture Models (GMM) results in higher recognition success.

The outline of the paper is the following. We first present the set of signal processing features used and propose new features that prove to be useful for instrument recognition. The feature selection strategy as well as the classification technique are then described. Finally, we proceed to the experimental study.

2. FEATURE EXTRACTION

Many features have been proposed for musical instrument recognition [4, 5, 1] describing various sound qualities. A number of these features become quite hard to extract robustly when dealing with musical phrases. Typically, note attack characteristics, although surely perceptually very important, are difficult to evaluate since onset detection is already intricate in our case [1]. Thus, a set of features which can be extracted in a more or less straightforward manner was chosen. In the following, we present a brief description of the features used. All of them are extracted on a frame basis.

2.1. Commonly used features

- **Temporal**. They consist of Autocorrelation Coefficients (AC) which were reported to be useful in [8],

[1] although there has been a number of proposals addressing this issue [7], there is no known system able to perform 100% successful onset detection due to the large variety of musical signals

in addition to Zero Crossing Rates (ZCR).

- **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 [5]. First, temporal amplitude envelopes were computed using a low-pass filtering of signal absolute complex envelopes, then a set of six coefficients was extracted as described in Eronen's work [5], namely AM frequency, AM strength and AM heuristic strength (for the two frequency ranges). Two coefficients were 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 were the product of tremolo frequency and tremolo strength, as well as the product of "graininess" frequency and "graininess" strength.
- **Cepstral**. Mel-Frequency Cepstral Coefficients (MFCC) are considered as well as their time first and second derivatives which are estimated over 9 successive frames.
- **Spectral**.

  Based on statistical moments. These included 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 [9] but also for musical instrument recognition [10]. They are denoted by Sx = Sc, Sw, Sa, Sf. Their time derivatives (δSx) (approximated over 9 successive frames) were also computed in order to provide us with an insight into spectral shape variation over time. It is worth to note that δSc can be seen as a quality of the vibrato playing technique since it embeds some frequency modulation information [5].

  MPEG7 spectrum flatness. A more precise description of the spectrum flatness was also used, namely MPEG-7 Audio Spectrum Flatness (ASF) [11] which is processed over a number of frequency bands. Indeed, this feature subset was found to be very useful for our task [10].

  Based on constant-Q transform. Frequency derivative of the constant-*Q* coefficients (describing spectral "irregularity" or "smoothness") were extracted as they were reported to be successful by Brown [2].

  Another useful feature consisted in a measure of the audio signal Frequency cutoff (Fc), also called frequency rolloff in some studies [12]. It was computed as the frequency below which 99% of the total spectrum energy was accounted.

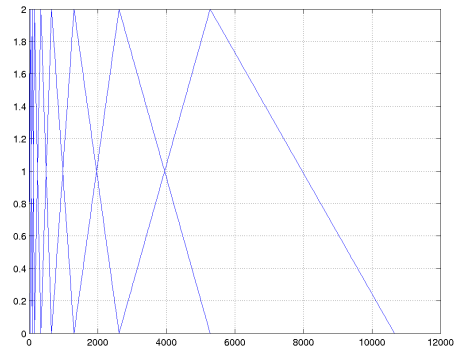

Figure 1. Octave band filterbank frequency response.

2.2. New features : Octave Band Signal Intensities

We introduce a new feature set which has been found very useful. The idea is to capture in a rough manner the harmonic structure of a musical sound, since it is desired to avoid 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. Figure 2.2 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 on 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.

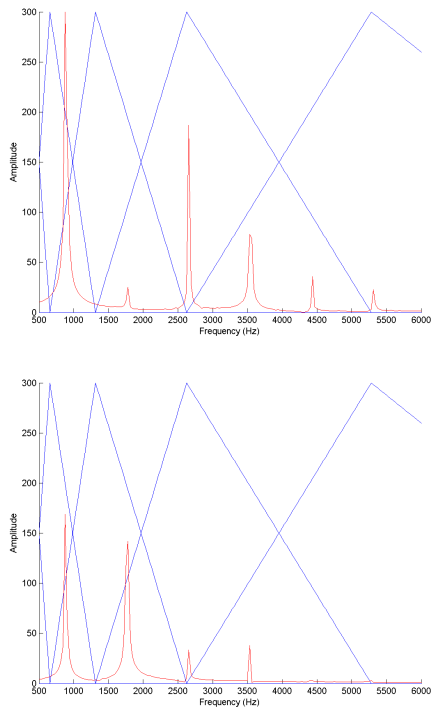

Figure 2. Amplitude spectrums of Alto Sax (top) and Bb Clarinet (bottom) playing the same note A4 and the octave band filterbank.

3. FEATURE SELECTION

Whenever an important number of candidate features are considered for a given classification task, it is very advantageous, not to say necessary to use feature selection techniques [3]. Such techniques aim at obtaining a "minimal" set of features which is the most efficient in discriminating between the classes under consideration, in the sense that selected features form the most informative and non-redundant subset of the original set of features. There has been a great deal of effort made to this end giving rise to a number of feature selection algorithms [3, 13]. We choose to use a technique proposed by Peeters [6] in the context of musical instrument classification. The author reported higher performance using the so-called "Inertia Ratio Maximization using Feature Space Projection" (IRMFSP) approach than the more classic "Correlation-based Feature Selection" (CFS) algorithm. Our main contribution here lies in adopting a pairwise feature selection strategy. The key idea is to select the subset of features that is the most efficient in discriminating between every possible pair of the considered instruments. We start by a brief description of the IRMFSP algorithm.

3.1. The IRMFSP algorithm

Feature selection is made iteratively with the aim to derive an optimal subset of d features amongst 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 ${n_k}^{\text{th}}$ 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 [14]) of the Between-class inertia B_i to the "average radius" of the scatter of all classes R_i defined as:

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

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.

In order to ensure the non-redundancy of the subset to choose, an orthogonalization step is introduced consecutive to every Inertia Ratio Maximization-based feature selection. At each iteration, ratio r_i maximization is performed yielding a new feature subset $\mathbf{X}_i$, and then 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 $\frac{r_d}{r_1} < \epsilon$ for a chosen ϵ, which means that the gain brought by the last selected feature has become non-significant.

3.2. Class pairwise feature selection

Our approach consists in performing the IRMFSP algorithm $\binom{K}{2}$ times [2], one processing for each pair of instruments (this will be referred to as $\binom{K}{2}$-IRMFSP by contrast to the classic approach denoted by 1-IRMFSP). A different set of features that is optimal in discriminating between two given instruments is searched for, in the perspective of a one vs one classification strategy. Hence, as many GMM classifiers as instrument pairs will be built based on different feature subsets. Beyond the improvement in recognition success (see section 5), the proposed scheme allows us to better understand instrument timbral differences. Indeed, it enables one to formulate statements 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 $\{i, j\}$. Additionally, it makes the analysis and optimization of classification performance more straightfor-

[2] $\binom{K}{2}$ being the number of combinations of 2 elements from K possible or the binomial coefficient

ward in the sense that it helps finding remedies to instrument confusions. For example, if the recognition success for a given instrument i is unsatisfactory because it is often confused with instrument j, it is reasonable to consider optimizing only the $\{i, j\}$ classifier.

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 [4, 5, 6]. The number of combinations to be considered at a time is then reduced to classes at the same level of taxonomy, rarely more than 4 classes.

4. CLASSIFICATION

4.1. The Gaussian Mixture Model (GMM)

The Gaussian Mixture model (GMM) has been widely used by the speech/speaker community since its introduction by Reynolds for text-independent speaker identification [15]. It was also successful for musical instrument recognition [2, 5]. In such a model, the distribution of the P-dimensional feature vectors is modeled by a Gaussian mixture density. For a given feature vector $\mathbf{x}$, the mixture density for instrument Ω_k is defined as :

$$p(\mathbf{x}|\Omega_k) = \sum_{i=1}^{M} w_i^k b_i^k(\mathbf{x}). \quad (2)$$

where the weighting factors w_i^k are positive scalars satisfying $\sum_{i=1}^{M} w_i^k = 1$. The density is then a weighted linear combination of M Gaussian component densities $b_i^k(\mathbf{x})$ with mean vector μ_i^k and covariance matrix $\mathbf{\Sigma}_i^k$ given by:

$$b_i^k(\mathbf{x}) = \frac{1}{(2\pi)^{P/2}|\Sigma_i^k|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(\mathbf{x}-\mu_i^k)'(\Sigma_i^k)^{-1}(\mathbf{x}-\mu_i^k)\right)} \quad (3)$$

The parameters of the model for the instrument k, denoted by $\lambda_k = \{w_i^k, \mu_i^k, \Sigma_i^k\}_{i=1,\ldots,M}$ are estimated thanks to the traditional Expectation-Maximization (EM) algorithm [16]. Classification is then usually made using the Maximum *a posteriori* Probability (MAP) decision rule. As an alternative, one can consider a one vs one decision strategy which can be very profitable as will be discussed in section 5. Classification is then performed using a "majority vote" rule applied over all possible class pairs and over L consecutive observations in time. For each pair of classes $\{\Omega_i, \Omega_j\}$, a positive vote is counted for the class Ω_i if

$$p(\mathbf{x}_t|\Omega_i) > p(\mathbf{x}_t|\Omega_j) \quad (4)$$

where $(p(\mathbf{x}_t|\Omega_k))_{k=i,j}$ is given in (2), $\mathbf{x}_t$ is the test feature vector observed at time t, and L is the total number of observations considered in taking decisions.

4.2. Rescaling and transforming the data

As a first pre-processing to GMM training, we introduce a rescaling stage which aims at homogenizing the highly varying dynamics of the different feature subsets. This is a well known technique in quantization problems whereby better precision is achieved by means of appropriate scale factors [17]. In our case, one scale factor is chosen for each feature subset in such a way that the resulting all-feature vectors have coefficients confined in the range $[0, 1]$.

The second pre-processing consists in using a Principal Component Analysis (PCA) transform in order to "denoise" the data [18]. The particularity of the approach rests on the fact that one PCA transform is computed for each instrument class (based on its training data). This has proven to be more efficient than a global PCA transform obtained from all-class data. PCA was performed as follows : for each instrument class, the covariance matrix of all related training feature vectors was computed and its Singular Value Decomposition (SVD) processed yielding

$$\mathbf{R_x} = \mathbf{UDV}^t,$$

where $\mathbf{R_x}$ is the covariance matrix, $\mathbf{U}$ and $\mathbf{V}$ are respectively the left and the right singular vector matrices, and $\mathbf{D}$ is the singular value matrix. The PCA transform matrix was then taken to be $\mathbf{W} = \mathbf{V}^t$ and classifiers were trained on the data $\mathbf{Y} = \mathbf{WX}$, where $\mathbf{X}$ is the matrix whose columns represent the training feature vectors. The same transform matrix $\mathbf{W}$ was applied on test feature vectors.

5. EXPERIMENTAL VALIDATION

Let us first give indications on various experimental parameters. The input signal was down-sampled to a 32-kHz sampling rate, it 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 16 ms for the extraction of all features except tremolo and roughness. Longer analysis length (960 ms and 480-ms hopsize) was used for the latter so as to measure the AM features properly. The AM feature values measured over each long window were then assigned to each 32-ms frame corresponding to the same time segment. All spectra were computed with a FFT after a Hamming window had been applied. 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.

The GMM was trained with ten iterations of the Expectation Maximization algorithm. Initialization consisted in clustering the observation space of accumulated feature vectors into $M = 16$ Voronoï regions thanks to the LBG

quantization procedure [19]. Initial means of the component densities were taken to be the centroids of the obtained clusters. Diagonal covariance matrices were used and initialized with empirical covariance coefficients of features from every Voronoï region.

Scoring was performed as follows : for each test signal, a decision regarding the instrument classification was taken every 0.47 s (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 0.47-s test segments.

5.1. Sound database for solo phrase recognition

Ten instruments were considered, namely, Alto Sax, Bassoon, Bb Clarinet, Flute, Oboe, Trumpet, French Horn, Violin, Cello and Piano. We used the same database that is described in [10] and presented in table 1. It is important to note that we used larger and more varied musical content than previous studies. This allowed us to achieve better training but also to draw statistically valid conclusions and assess the generalization capabilities of our classification scheme.

	Train	Srcs	Trks	nTests	Test
Alto Sax	9.37	10	19	682	5.46
Bassoon	3.33	5	9	287	2.30
Bb Clarinet	13.13	10	26	1077	8.62
Flute	17.74	8	24	2173	17.38
Oboe	18.29	8	24	2162	17.30
French Horn	4.61	5	13	369	2.95
Trumpet	20.14	9	73	2399	19.19
Cello	19.26	7	20	2332	18.66
Violin	22.67	11	31	2447	19.58
Piano	20.48	8	15	1862	14.90

Table 1. Sound database - *Srcs* is the total number of distinct sources used during test; *Trks* is the total number of tracks from CDs during test; *nTests* is the number of tests performed (1 test = 1 class decision over 0.47 s); *Total train* and *Total test* are the total durations of respectively train and test material in minutes.

5.2. Features

Table 2 sums up the feature subsets used together with the features selected in the 1-IRMFSP configuration with a stop criterion $\epsilon = 10^{-5}$. A total of 19 features were selected including MFCC, Sx, ASF, OBSI and OBSIR. Not only were OBSI attributes selected in priority but also they are the feature subset that is the most largely represented in the set of selected features. Since the most relevant features are selected in decreasing order of importance by the IRMFSP algorithm, it can be deduced that these attributes are useful for the instrument recognition task. Only the 4 first MFCCs were selected which is quite a low number compared to the 10 or 12 coefficients usually used for sound source recognition.

Feature subset	Size	Selected
AC=[A1,...,A49]	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 2. Feature subsets and 1-IRMFSP results

Using the same stop criterion $\binom{10}{2}$-IRMFSP was performed (for the 45 possible pair combinations) yielding an average number of 19 features per instrument pair. The number of selected features varied from 9 (for the Piano/Violin confrontation) to 44 (for Bb Clarinet versus Flute). This is another benefit of the chosen methodology : features are specifically tuned to the context, whenever two instruments are easily distinguished, the number of needed features is smaller. Examples of class pairwise feature selection results are presented in table 3.
One can draw the following conclusions.

- Some features were never selected; this is the case for the first and second time derivatives of Spectral centroid (δSc, δ^2Sc), Spectral width (δSw, δ^2Sw), Spectral asymmetry (δSa, δ^2Sa) and Spectral flatness (δSf, δ^2Sf). Also, cepstrum first time derivatives (except δC0) and second time derivatives (except δ^2C0) and the product of "graininess" frequency and "graininess" strength.

- Sc, Sw, Sa and Sf together with MPEG-7 ASF coefficients and OBSI/OBSIR were the most successful features since an important subset of them was always selected for every instrument pair. It is worth to note that Sc was not considered useful in a number of cases, probably because the same information was embedded in other features describing the spectral shape. The average number of selected MFCCs was 4 (consistent with the 1-IRMFSP findings).

- Some other features, although not selected very often, were useful in particular situations. In fact, Spectral "irregularity" coefficients (Si) were considered particularly useful for combinations involving the Bb Clarinet 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 4 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.

Bb Clarinet/Alto Sax	Bb Clarinet/Bassoon	Bb Clarinet/Flute	Bb Clarinet/French Horn
C1,..,C3,C6,...,C8,C11 Sc,Sw,Sa,Sf A16,A22 S12 OR5	C1,...,C4 Sc,Sw,Sa A21,...,A23 S12,S18 O5,...,O7	R5,R10,R23,R42-ZCR C1,..,C3,C6-δ^2C0 Sc,Sw,Sa,Sf A5,A9,A10,A18,A20,A22,A23 S7,S8,S15,S16,S18,S19 Fc-O1,..,O8-OR1,...,OR7 AM5	ZCR-C1,..,C6 Sc,Sw,Sa,Sf A2,A3,A5,A6,A9,A10,A14,A18,A20,A23 S9,S13,S14,S15,S16,S20 Fc-OR5,OR6 AM1,AM2,AM3,AM6

Bb Clarinet/Trumpet	Bb Clarinet/Cello	Bb Clarinet/Violin	Bb Clarinet/Piano	Bb Clarinet/Oboe
A8-C2,C3 Sw,Sa,Sf S15,S16,S19 O1,O5,O6,O7 OR5,OR7	A1-C2,C3 Sw,Sa,Sf A22-S19 O5,O9 AM1	C1,...,C3 Sw,Sa,Sf A20,A22,A23 Fc O4,O5	A1-C1,...,C4 Sw,Sa C2,..,C5,C7 A13,A18,A20,A22,A23 Fc-O2,O6,O7,O8 OR6,OR7	A1,A8,A18 Sc,Sw,Sa-ASF22 S11,S14-O2,O4,O6,...,O8 OR5,OR7

Table 3. Features selected by the $\binom{10}{2}$-IRMFSP algorithm for a few examples.

5.3. Classification results

For the one vs one scheme, 45 GMMs were trained based on features selected for each combination of 2 instruments through a $\binom{10}{2}$-IRMFSP approach. No particular strategy was employed at the test stage to cope with the indeterminations resulting from the use of the majority vote rule, typically the cases where two classes received the same number of votes. Under these circumstances, we merely adopted random selection in deciding the class for the related test segment. It is worth to note that better performance could be achieved with the one vs one scheme using more sophisticated techniques in coupling the pairwise decisions in order to decide the class associated with a given test segment (see [20] for example).
In parallel, one GMM per instrument was trained using the data obtained thanks to a classic 1-IRMFSP and MAP decision was used at the test stage.

% correct	oVo-nr	oVo	oVo+PCA	MAP
Alto Sax	58.03	53.48	56.36	65.76
Bassoon	60.14	64.86	61.23	63.41
Bb Clarinet	63.78	80.54	84.49	71.48
Flute	56.69	86.08	84.79	75.70
Oboe	82.29	81.87	82.58	75.31
French Horn	62.78	65.96	71.49	57.45
Trumpet	71.55	64.02	68.19	79.70
Cello	90.47	89.51	90.88	89.14
Violin	95.53	95.48	94.43	88.61
Piano	92.82	96.41	96.35	72.21
Average	73,41	77,82	79,08	73,88

Table 4. Instrument recognition success with different classification schemes - *oVo* stands for one vs one, *nr* stands for no rescaling

First, the benefit of the rescaling procedure is highlighted in columns 2 and 3 of table 4, presenting the recognition success with the one vs one approach when features were rescaled (column 3) and without such pre-processing (column 2). The average improvement with rescaling is 4.41%. Very significant improvement is achieved especially for the Bb Clarinet (+16.7%) and the Flute (+29.4%). Indeed, rescaling seems to help the EM optimization in estimating the means of the mixture density-components as the data can be quantized more precisely at the initialization stage. Even though a decline in performance is observed in some cases (eg. Trumpet) it remains quite small compared to the improvement and could probably be avoided with a better optimization of the scale factors.

Results obtained with PCA-transformed data are given in column 4. Some improvement is observed for the Alto Sax, the Oboe and the French Horn. The success rates remain hardly changed for the rest. It appears that, since irrelevant features are removed thanks to the feature selection algorithm, the use of PCA becomes less effective compared to the case where no feature selection techniques are exploited.

Let us now compare our classification scheme to classification using 1-IRMFSP in association with a classic GMM approach (with as many models as instruments) and the MAP criterion (column 5). Our proposal performs substantially better in almost all cases. The success rate reaches 96.41 % for the Piano with the one vs one approach while it is only 72.21 % with the classic methodology (+24.2%). However, there are two exceptions, namely Alto Sax and Trumpet for which it is believed that rescaling and model training was sub-optimal (see column 1 and 2). The Alto Sax was confused with the Violin 37% of the time and the Trumpet with the Oboe 11% of the time and with the Flute 9% of the time. A great advantage of the one vs one approach resides in the fact that one could consider optimizing only the Alto Sax/Violin classifier in order to improve the Alto Sax recognition rate. The optimization should be concerned with both features and classification techniques. One should focus on finding new specific descriptors that could be amenable to better discrimination between Alto Sax and Violin, but also on more adapted classifiers. In fact, it is also possible to consider different classifiers for each instrument pair using for example the best of GMM and Support Vector Machines [21] with kernels specifically tuned to a certain combina-

tion of 2 instruments. This can yield more effectiveness in recognition.

The last experiment consisted in modifying the IRMFSP stop criterion by choosing a smaller ϵ in order to check the effect of selecting more features on the recognition task success. A value of $\epsilon = 10^{-6}$ yielded 33 features with 1-IRMFSP and an average number of 38 features when using the $\binom{10}{2}$-IRMFSP. Results are given in table 5. Scores are slightly higher and the one vs one approach remains in the overall performance more efficient.

% correct	MAP	oVo
Alto Sax	71.82	55.15
Bassoon	61.96	68.84
Bb Clarinet	81.12	84.87
Flute	82.28	88.87
Oboe	77.83	85.07
French Horn	63.83	71.91
Trumpet	82.44	71.77
Cello	91.30	87.81
Violin	94.43	96.64
Piano	90.80	98.17
Average	79,78	80,91

Table 5. Instrument recognition success with $\epsilon = 10^{-6}$, 33 features for 1-IRMFSP and an average of 38 features for $\binom{10}{2}$-IRMFSP.

6. CONCLUSION

In this work, a one vs one classification scheme was proposed for the recognition of 10 musical instruments on real solo performance. A high number of signal processing features was considered including new proposals that have proven to be successful for this task. Moreover, it has been shown that it is advantageous to tackle the feature selection problem in a pairwise fashion whereby the most relevant features in discriminating between two given instruments are found. The chosen approach entails higher recognition success and allows us to analyze the recognition system performance and look for enhancements in a more straightforward manner.

Additionally, a data rescaling procedure has been introduced that lead to substantial improvement in classification performance.

Future work will consider classifiers specifically adapted to every instrument pair, particularly, Support Vector Machines in a scheme where the best kernel is selected for a given combination of 2 instruments.

7. REFERENCES

[1] P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *New Music Research*, 32.1, 2003.

[2] Judith C. Brown, Olivier Houix, and Stephen McAdams. Feature dependence in the automatic identification of musical woodwind instruments. *Journal of the Acoustical Society of America*, 109(3):1064–1072, March 2000.

[3] Huan Liu and Hiroshi Motoda. *Feature selection for data mining and knowledge discovery*. Kluwer academic publishers, 1998.

[4] Keith Dana Martin. *Sound-Source Recognition : A Theory and Computational Model*. PhD thesis, Massachusets Institue of Technology, June 1999.

[5] Antti Eronen. Automatic musical instrument recognition. Master's thesis, Tampere University of Technology, April 2001.

[6] Geoffroy Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *AES 115th convention, New York, USA*, October 2003.

[7] Juan P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler. A tutorial on onset detection in music signals. *IEEE trans. on speech and audio processing*, to be published.

[8] J. Brown. Musical instrument identification using autocorrelation coefficients. In *International Symposium on Musical Acoustics*, pages 291–295, 1998.

[9] Olivier Gillet and Gaël Richard. Automatic transcription of drum loops. In *IEEE ICASSP, Montréal, Canada*, May 2004.

[10] Slim Essid, Gaël Richard, and Bertrand David. Efficient musical instrument recognition on solo performance using basic features. In *AES 25th international conference, London, United Kingdom*, June 2004.

[11] Information technology - multimedia content description interface - part 4: Audio, June 2001. ISO/IEC FDIS 15938-4:2001(E).

[12] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM, 2004.

[13] L. Molina, L. Belanche, and A. Nebot. Feature selection algorithms :a survey and experimental evaluation. In *International conference on data mining, Maebashi City, Japan*, 2002.

[14] P. E. Hart Richard Duda. *Pattern Classification and Scence Analysis*. Wiley-Interscience. John Wiley Sons, 1973.

[15] D. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.

[16] Todd K. Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, pages 47–60, nov 1996.

[17] A. Gersho and R. Gray. *Vector quantization and signal compression*. Kluwer academic publishers, 1992.

[18] Slim Essid, Gaël Richard, and Bertrand David. Musical instrument recognition on solo performance. In *EUSIPCO, Vienna, Austria*, September 2004.

[19] Y. Lindo, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. In *IEEE Trans. Communication*, pages 84–95, January 1980.

[20] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. Technical report, Stanford University and university of Toronto, 1996.

[21] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Journal of Data Mining and knowledge Discovery*, 2(2):1–43, 1998.