Efficient musical instrument recognition on solo performance music using basic features

Slim Essid, Gael Richard, Bertrand David

Efficient musical instrument recognition on solo performance music using basic features

Slim Essid[1], Gaël Richard[1], and Bertrand David[1]

[1] *GET-ENST (Télécom Paris), 37-39 rue Dareau, 75014 Paris, France*

Correspondence should be addressed to Slim Essid (`Slim.Essid@enst.fr`)

ABSTRACT

Musical instrument recognition has gained growing concern for the promise it holds towards advances in musical content description. The present study pursues the goal of showing the efficiency of some basic features for such a recognition task in the realistic situation where solo musical phrases are played. A large and varied database of sounds assembled from different commercial recordings is used to ensure better training and testing conditions, in terms of statistical efficiency. It is found that when combining cepstral features with others describing the audio signal spectral shape, a high recognition accuracy can be achieved in association with Support Vector Machine classification when using a Radial Basis Function kernel.

1. INTRODUCTION

The ability to recognize musical instruments from real world musical performance stands as a key capability of an efficient audio indexing system. It is of great importance for applications such as content based search or automatic extraction of musical scores. Although this issue received a growing interest, the majority of studies only considered isolated musical notes as input to the recognition system [1, 2]. In choosing to process solo musical phrases from commercial Compact Discs (CD), we are putting our study in the line of Brown's [3], Martin's [4] and Marques' [5] work. In fact, addressing a musical content from real world solo performance seems to be the most promising approach for immediate applications provided that instrument recognition in polyphonic music context, *i.e* involving more than one instrument at a time, remains a complex problem which has been barely addressed.

A major issue in building efficient instrument recognition systems is the choice of proper signal processing features likely to result in effective discrimination between the different instruments when recurring to more or less elaborate classification techniques. While a great deal of effort has been dedicated to this end, giving rise to a large number of potentially useful features [6, 4, 1, 3, 7], only a few proposals can be retained in the context of musical phrases, since their processing may become quite intricate when concurrent notes are played. The purpose of this work is thus to study the effect of combining simple and robust features on the efficiency of the instrument recognition system.

Furthermore, a number of studies relied on limited sound databases both in size and diversity which, prevented from achieving efficient model training but also from drawing statistically valid conclusions. The use of a much larger sound database of excerpts from many different recording conditions, with different instrument instances and performers is thus an important aspect of this work.

The outline of the paper is the following. First, we introduce the set of features which were chosen. Second, we present a brief description of Support Vector Machine (SVM) classification which was exploited in our work. Finally, we proceed to the experimental study that was conducted on the efficiency of the proposed features leading to high recognition accuracies in association with Principal Component Analysis (PCA) and SVM classification with a Radial Basis Function (RBF) kernel.

2. FEATURE EXTRACTION

We chose for this study, to focus on a reduced number of simple, yet robust features, which can be extracted in a straightforward manner and still result in satisfactory recognition success rate. This means that features related to audio signal pitch and attack characteristics were avoided (typically onset duration and slope, harmonic structure, tristimulus, etc.). In effect, the underlying extraction stages, namely multi-pitch estimation and onset detection, give rise to problems that remain partially unsolved whenever concurrent notes are played, given the state of the art. Another benefit of using fewer features is the resulting reduction of classification algorithms computational cost since the dimensionality of the feature space is smaller. Mel-Frequency Cepstral Coefficients (MFCC) have proven successful for various sound source classification tasks including instrument classification [3, 1], thus, they were used as baseline features. Our approach then consisted in appending other basic features that could improve instrument discrimination. These were time derivative of MFCC (which will be referred to as ΔMFCC) as well as a set of features describing the audio signal spectral shape which have proven successful for drum loop transcription [8]. They are obtained from the statistical moments μ_i and defined as follows :

- the spectral centroid, $S_c = \mu_1$;
- the spectral width, $S_w = \sqrt{\mu_2 - \mu_1^2}$;
- the spectral asymmetry defined from the spectral skewness, $S_a = \frac{2(\mu_1)^3 - 3\mu_1\mu_2 + \mu_3}{S_w^3}$;
- the spectral flatness defined from the spectral kurtosis, $S_f = \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{S_w^4} - 3$;

where $\mu_i = \frac{\sum_{k=0}^{N-1} f_k^i A_k}{\sum_{k=0}^{N-1} A_k}$, with A_k the amplitude of the k-th frequency component f_k of the input signal Fourier transform. Additionally, an alternative description of the spectrum flatness was used, namely MPEG-7 Audio Spectrum Flatness (ASF) [6] which is processed over a number of frequency bands. This was not used in previous work on instrument recognition, yet it was found quite useful as will be discussed later in the paper.

3. SUPPORT VECTOR MACHINES

The classification approach used in this study is known as Support Vectors Machines (SVM) which have been used successfully for various classification tasks. Considering two classes, SVM try to find the hyperplane that separates the features related to each class with the best possible margin. In the case where the data is non-linearly separable, SVM map the P-dimensional input feature space into a higher dimension space where the two classes become linearly separable, thanks to a Kernel function $K(\mathbf{x}, \mathbf{y})$ such that

$$K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}).\Phi(\mathbf{y}),$$

where $\Phi : \mathrm{R}^P \longmapsto \mathrm{H}$ is a map to the high dimension space H. SVM classification is very advantageous in the sense that it has interesting generalization properties. Interested readers are referred to [9] for detailed description and discussion of SVM.
Such classifiers can perform binary classification and regression tasks. It can also be adapted to perform N-class classification. To this end, we adopted the "one Vs one" strategy which consists in building one SVM per possible combination of two instruments. Classification is then performed using a "majority vote" rule applied over all possible pairs and over a number of consecutive observations in time, *i.e* the algorithm selects the class with the largest positive outputs over a subset of feature vectors of a test signal.

4. EXPERIMENTAL STUDY

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. 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. As far as cepstral features are concerned, the ten first MFCCs (not including the zero-th coefficient) and the ten first ΔMFCCs were selected. The ASF

features included 23 coefficients.
Scoring was performed as follows : for each test signal, a decision regarding the instrument class it belonged to 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.

4.1. 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. 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 used in our experiments with regard to recording conditions, performers and instrument instances. Sound samples were excerpted from 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 Télécom Paris studio. The selection of recording excerpts used in the training set was made randomly under the constraint that at least 15 minutes of data were assembled. Whenever this was not possible, at least 2 minutes 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 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 1 sums up the properties of the data used in our experiments. Let us emphasize that we used much larger and more varied musical content than previous studies[1] allowing us to achieve better training but also to draw statistically valid conclusions and assess the generalization capabilities of our classification scheme.

[1] the average length of sound data was rather around 4 minutes and the average number of sources was rather 4 for each instrument

4.2. On features

The first study is meant to highlight the contribution of various feature subsets to instrument recognition success. Basic linear SVM classification is here used since the focus is put, at this stage, on the importance of features. Thus, recognition experiments were undertaken using first, feature vectors composed of only MFCCs, second, feature vectors composed of both MFCCs and ΔMFCCS, third appending S_c, S_w, S_a and S_f (which will be referred to as S_x) to the previous features and finally appending ASF to form a 47-dimension feature space (with 10 MFCCs, 10 ΔMFCCs, 4 S_x and 23 ASF coefficients). This ordering in extending the feature vectors is motivated by the discussions on features in previous work, in such a way that the most "popular" features were taken in priority. The Obtained recognition accuracy for the ten considered instruments is given in the four first columns of table 2.
Using the baseline features, *i.e* MFCC, high recognition accuracy can be achieved for some instruments such as the Piano (92.2 %), the Cello (87.2 %) and the Oboe (79.2 %) while unacceptable performance is found for the Bb Clarinet (41.0 %) and even worse for the French Horn which is very scarcely successfully identified (chance selection would have worked better in this case). As far as class confusions are concerned, it is worth noting that under these conditions, the French Horn is identified as Piano with a rate of 44.1 % and as Bassoon with a rate of 37.1 %, while the Bb Clarinet is confused with the Piano in 20.8 % of the tests and with the Flute in 18.3 % of the tests. This poor performance for the Bb clarinet is however not really surprising since this instrument is characterized by the prominence of its odd harmonics. Clearly, a feature measuring the ratio of odd and even harmonics would be particularly useful for this case but such a parameter is more difficult to estimate on real solo performance and would be nearly impossible to estimate on sound mixtures. Appending ΔMFCC features enables the classification algorithm to better discriminate the Bb Clarinet from the Piano as it is then confused with the latter in only 9 % of the cases, which results in better recognition accuracy (56.3 %) even though it is still as much confused with the Flute as with only MFCC (18.8 %). In fact, consistent with the findings of Brown [10], ΔMFCCs have been found inefficient for discriminating between instru-

	Total train (mn)	Sources	Tracks	Nb tests	Total test (mn)
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 - *Sources* is the total number of distinct sources used during test; *Tracks* is the total number of tracks from CDs during test; *Nb tests* 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.

ments from the woodwind family, which is pointed out by the class confusion rates (not presented here for lack of space). Significant improvement is also achieved for the recognition of the Violin (plus 5 points). However, for the majority of the considered instruments the recognition success is either hardly changed or smaller, especially for the Bassoon (minus 15 points). The French Horn is less often confused with the Bassoon (15.7 %) yet it is more frequently identified as Piano with a rate of 68.0 %. Using the combination of MFCC, ΔMFCC and S_x as features results in better performance for the recognition of the Bassoon, the Bb Clarinet, the Violin and the French Horn. Finally, classification based on all proposed features results in overall important improvement in recognition accuracy. ASF turns out to be a useful feature for instrument recognition although it should be computed in a more compact form reducing the number of output coefficients which is here quite high (23) compared to other feature subset sizes (only 10 MFCCs).

In a nutshell, when extending the set of features, while in some cases the new features will help the classification algorithm to better discriminate a given instrument from all others (as for the Violin), in some other cases, these new features will bring more confusion, eventually with only a subset of possible instrument classes, even though the complete set of features leads to significant improvement compared to the baseline features. It is thought that an approach holding much promise would consider feature selection techniques for instruments taken pairwise (this is totally compatible with our "One Vs One" classification strategy) in order to find out which features are the most efficient in discriminating between any pair of instruments.

%	C	+ΔC	+S_x	+ASF	PCA	RBF
AltoSax	62.5	58.2	60.9	66.1	65.1	
Bassoon	51.2	36.6	52.0	50.2	50.2	
Clarinet	41.0	56.3	61.5	77.4	76.0	
Flute	80.7	81.0	79.7	86.8	87.2	
Oboe	79.2	78.1	73.3	74.6	75.4	
Fr Horn	0.0	0.8	24.1	54.0	54.7	
Trumpet	79.0	78.2	76.6	81.1	81.4	
Cello	87.2	88.1	82.9	86.7	86.4	
Violin	77.5	82.7	90.4	92.2	88.7	
Piano	92.2	92.8	86.2	93.5	93.1	

Table 2: Obtained recognition accuracies. C stands for MFCC and ΔC for ΔMFCC - '+' means appending features

4.3. Improving the classification scheme

The next experiments were concerned with using more elaborate classification tools in order to improve the overall recognition accuracy. First, Principal Component Analysis (PCA) [11] was considered in order to "de-noise" the feature space and reduce the dimensionality of the problem. The "de-noising" effect is due to the fact that the most relevant information gets concentrated in the first few components of the transformed feature vectors which correspond to directions of maximum energy. Recognition ac-

curacies obtained with linear SVM classification operating on PCA transformed data into a reduced 35-dimension space are given in column 5 of table 2. For all instruments but the Violin, the success rate remains quite unchanged while the dimensionality of the problem has been reduced from 47 to 35, which is advantageous as far as computational cost is concerned. However, no significant improvement in recognition success was found when using SVM classification compared to the case where PCA is associated with Gaussian Mixture Model (GMM) based classification [12].
Finally, we performed SVM classification with a Radial Basis Function kernel of the form

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

with $\gamma = 1$. The obtained results are given in column 6 of table 2. High recognition accuracies are found with an average of ?? %. Note that very short time decisions are taken (one decision every 0.47 s). Higher accuracy could be reached using longer term decisions yet it would prevent real-time applications which can still be considered with the chosen decision length.

5. CONCLUSION

We have proposed a study on the efficiency of simple features (that can be computed robustly) for instrument recognition systems in the context of solo musical performance. Experimental sound material was large enough to allow us to perform proper training of SVM classifiers as well as to assess the statistical validity of our conclusions. It has been shown that the combination of cepstral coefficients with features describing the audio signal spectral shape results in high accuracy recognition for instruments belonging to the different families even over short-term decision lengths. Future work will consider the selection of the most relevant features in discriminating between any pair of instruments as well as the extension of the present feature set.

6. REFERENCES

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

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

[3] 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, mar 2000.

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

[5] Janet Marques and Pedro J. Moreno. A study of musical instrument classification using gaussian mixture models and support vector machines. Technical report, 1999.

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

[7] Shlomo Dubnov and Xavier Rodet. Timbre recognition with combined stationary and temporal features. In *International Computer Music Conference*, 1998.

[8] Olivier Gillet and Gaël Richard. Automatic transcription of drum loops. In *IEEE ICASSP*, mai 2004.

[9] 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.

[10] Judith C. Brown. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *Journal of the Acoustical Society of America*, 105(3):1933–1941, mar 1999.

[11] M. Partridge and M. Jabri. Robust principal component analysis. In *IEEE Signal Processing Society Workshop*, pages 289–298, dec 2000.

[12] Slim Essid, Gaël Richard, and Bertrand David. Musical instrument recognition on solo performance. *to be published*, 2004.