MUSICAL INSTRUMENT RECOGNITION ON SOLO PERFORMANCES

Slim Essid, Gaël Richard, Bertrand David

MUSICAL INSTRUMENT RECOGNITION ON SOLO PERFORMANCES

Slim ESSID, Gaël RICHARD, Bertrand DAVID

GET - ENST (Télécom Paris)
46, rue Barrault - 75634 Paris Cedex 13 - FRANCE
phone: +33 1 45 81 71 71, e-mail: Slim.Essid, Gael.Richard, Bertrand.David@enst.fr

ABSTRACT

Musical instrument recognition is one of the important goals of musical signal indexing. If much effort has already been dedicated to the automatic recognition of musical instruments, most studies were based on limited amounts of data which often included only isolated notes. In this paper, two statistical approaches, namely the Gaussian Mixture Model (GMM) and the Support Vector Machines (SVM), are studied for the recognition of woodwind instruments using a large database of isolated notes and solo excerpts extracted from many different sources. Furthermore, it is shown that the use of Principal Component Analysis (PCA) to transform the feature data significantly increases the recognition accuracy. The recognition rates obtained range from 52.0 % for Bb Clarinet up to 96.0 % for Oboe.

1. INTRODUCTION

Musical instrument recognition has gained more and more interest as the need for multimedia description tools has become obvious in the lights of the MPEG-7 standardization effort [3]. As far as musical content is concerned, the challenge is to reach configurations where complex mixtures of sound could be fully labeled (for example, in terms of mode, style and rhythm,...) and furthermore indexed in terms of musical events in order to permit, at a very high level of description, the extraction of a score-type representation. One could then be able to formulate requests such as "find Jazz trumpet solo parts played in $C^{\#}$ in the middle of a musical database".

The task is very complex and many problems remain unsolved given the current state of the art. For instance, very few attempts have been made on a musical content involving more than one instrument playing at a time. Previous work on musical instrument recognition mainly focused on the case where isolated notes were played motivated by the hypothesis that separation of the different sound sources in the signal, followed by note segmentation could be achieved in a first stage of processing (see [8] for a complete review). Yet, the task of source separation and segmentation can be even more intricate than source recognition. It is thought that the most promising approach which moreover could give rise to immediate applications, is to consider the recognition of solo musical phrases taken from commercial recordings.

To our knowledge, there was only three studies reporting significant performance that adopted these conditions, by Brown *et al.* [4], Martin [10] and Marques [9]. Both parametric and non-parametric classification techniques were used. For instance, encouraging results have been found with Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) which have proven successful in various classification tasks, particularly in speech recognition and speaker identification. Unfortunately, in contrast to the speech/speaker community, there has been no specific common sound database of musical solo excerpts of appropriate size and with enough diversity that could have been used for evaluating the relative performance of the proposed musical instrument recognition systems.

In this work, two statistical approaches, namely the Gaussian Mixture Model (GMM) and the Support Vector Machines (SVM), are studied for the recognition of woodwind instruments using a large database of isolated notes and solo excerpts extracted from many different sources. Our work extends Marques study [9] in using much larger sound databases for both training and testing and in introducing Principal Component Analysis (PCA) to "de-noise" the feature space.

The use of a much larger sound database of excerpts recorded in different conditions, with different instrument instances and performers clearly permits to better grasp the variability of realistic situations. The use of such databases is essential to build classifiers with better recognition accuracy and ability to cope with data variability.

The paper is organized as follows. In the first place, using Mel-Frequency Cepstrum Coefficients (MFCC) as features, a study on isolated notes is presented to test several variations on the classification strategies for model training and decision rules. Principal Component Analysis (PCA) is then introduced and its use as preprocessing to classification is discussed. Finally, GMM and SVM classification are used on PCA transformed data from recordings and their performance is compared.

2. FEATURE EXTRACTION

Many features have been proposed for musical instrument recognition [8] and applied with a certain success in the case of classification on isolated notes. Nevertheless, many of these features cannot be extracted in a robust manner, especially when dealing with phrases from real musical performances. For example, features related to sound attack, harmonic structure, vibrato and tremolo, etc., have been found to be very useful [10, 6]. Yet, the underlying extraction stages, namely, onset detection and multi-pitch estimation give rise to problems that remain partially unsolved whenever concurrent notes are played, given the state of the art.

The Mel-Frequency Cepstral Coefficients (MFCC) have proven successful for our task [4, 6] and have the advantage of being easily extracted, therefore they were chosen as features for this study. Delta Cepstrum is not used, since it was found useless for the woodwind instrument recognition, consistent with the findings of Brown [4]. It is important to note that other useful features (related to spectral shape, for example) could be added, but the primary goal of this paper is to assess the performance of the classification techniques on a large sound database considering baseline features

such as MFCCs.

The input signal was down-sampled to a 32 kHz sampling rate, centered with respect to its temporal mean and its amplitude 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. The cepstrum was 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 and then discarded. A feature vector consisted of the 10 first cepstral coefficients not including the zeroth coefficient.

3. THEORETICAL BACKGROUND ON CLASSIFICATION

3.1 The Gaussian Mixture Model

The Gaussian Mixture model (GMM) has been widely used by the speech/speaker community since its introduction by Reynolds for text-independent speaker identification [14]. It was also successful for musical instrument recognition [9, 4, 6]. In such a model, the distribution of feature vectors (in our case the feature vectors of P MFCCs, with $P = 10$ for a given instrument class) are modeled by a Gaussian mixture density. For a given feature vector $\mathbf{x}$, the mixture density for instrument Ω_k is defined as :

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

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:

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

The parameters of the model for the instrument k, denoted by $\lambda_k = \{p_i^k, \mu_i^k, \Sigma_i^k\}_{i=1,\ldots,M}$ are estimated thanks to the traditional Expectation-Maximization (EM) algorithm [11]. Classification is then usually made by using the Maximum *a posteriori* Probability (MAP) decision rule, which thanks to Baye's rule, can be written as

$$\hat{\Omega} = \arg\max_{1\leq k\leq N} \sum_{t=1}^{T} \log p(\mathbf{x}_t|\Omega_k) \quad (3)$$

where N is the number of instrument classes, $p(\mathbf{x}_t|\Omega_k)$ is given in (1), $\mathbf{x}_t$ is the test feature vector observed at time t, and T is the total number of observations considered in taking decisions.

3.2 Support Vector Machines

The other classification approach used in this study is known as Support Vectors Machines (SVM) which have been used 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. Such classifiers can perform binary classification and regression estimation tasks but can also be adapted to perform N-class classification. SVM classification is very advantageous in the sensse that it has interesting generalization properties. Interested readers are referred to [5] for detailed description and discussion of SVM.

4. EXPERIMENTAL STUDY

4.1 Sound database for isolated note recognition

Three musical note collections were used for a preliminary study, namely, McGill University Master Samples (MUMS)[12], IRCAM Studio Online collection (SOL) [1] and the University of Iowa Electronic Music Studios samples [2]. The aim of this study was to investigate a number of variations on the classification techniques with reasonable overall computational cost and provide feedback to reference work on isolated note recognition. Leave-one-out cross-validation was used, in the sense that two out of the three collections were used as training set and the remaining collection was used as test set with the three possible combinations.

4.2 Sound database for solo phrase recognition

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, Jazz music or educative material for music teaching.

The task has been particularly difficult and time-consuming with some instruments for which unaccompanied solo performances may be very uncommon, as it is the case for the Bassoon. Additionally, Sax and Bb Clarinet solo phrases performed by two amateur players were recorded at Télécom Paris studio. Although this study is limited to five woodwind instruments, it is conducted on a much larger and more varied database than previous studies allowing to assess the generalization properties of the classification task. It was thought that it would be more interesting to consider fewer instruments with enough training and test data than many instruments with insufficient samples. All note collections were then included in the training set for the solo phrases recognition experiments in addition to selected recording samples.

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. Ideally, never would the same CD provide excerpts for both training and test sets, but, in some cases, it hasn't been 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 any tracks that were used in the training set. All samples not used for training were tested so as to provide tight confidence ranges on the success rates. Table 1 presents an overview of the sound database and training/test division. The amounts of data used in previous work on instrument recognition on solo phrases are summed up in table 2. Note that much more data is used in our experiments.

4.3 Study on musical note collections

4.3.1 GMM classification

The GMM was trained with ten iterations of the Expectation Maximization algorithm. Initialization consisted in clustering the observation space of accumulated MFCC vectors into M Voronoï regions thanks to the LBG quantization procedure. 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 MFCCs from

	Sources	*Tracks*	*Train* (mn)	*Test* (mn)
Alto Sax	12	150	31.71	1.45
Bassoon	11	115	16.73	1.41
Bb Clarinet	16	155	21.67	7.12
Flute	13	216	27.86	141.49
Oboe	12	133	20.83	52.88

Table 1: Sound database - *Sources* is the number of distinct sources used; *Tracks* is the number of tracks from CDs and files from collections; *Train*, respectively *Test*, is the total length of the training data, respectively test data, in minutes.

	Sources	*Train* (mn)	*Test* (mn)
Brown [4]	na	0.9-5.5	1.0-4.0
Martin [10]	2-9	0.23-35.5	0.9-35.5
Marques [9]	2-2	3.4-3.4	0.3-0.3

Table 2: Sound database - *Sources* is the number of distinct sources used; *Train*, respectively *Test*, is the total length of the training data, respectively test data, in minutes, minimum and maximum durations are given. na stands for not available.

every Voronoï region. The number of components in the mixture was varied from 4 to 16. $M = 16$ gave the best results. Recognition was made by taking decisions every 0.469 s corresponding to the accumulation of $T = 30$ observation frames.

Three decision rules were considered depending on the chosen classification strategy, namely MAP decision, "one Vs one" and "one Vs all". Note that, when considering the "one Vs all" approach, one must set up two models for each class Ω_k, with $k = 1,...,5$: one model λ_k for Ω_k, but also one model, $\bar{\lambda}_k$ for the data representing all classes but Ω_k, referred to as $\bar{\Omega}_k$. For the "one Vs all" approach, a given test frame $\mathbf{x}_t$ was classified as belonging to class Ω_k if $\text{Prob}(\mathbf{x}_t|\lambda_k) \geq \text{Prob}(\mathbf{x}_t|\bar{\lambda}_k)$ and the class with the largest positive outputs over T observations was selected. For the "one Vs one" case, a "majority vote" rule was applied over all possible pairs and over T observations.

Table 3 shows the recognition results obtained with leave-one-out cross-validation for the note collections. The "one Vs one" approach yielded results very close to those obtained with the "one Vs all" approach, thus they are not presented here. "one Vs all" performed slightly better than MAP but the improvement remains quite small in most cases. Nevertheless, "one Vs all" can be more advantageous in a scheme where instrument-specific parametrisation would be aimed at. For example, one could use feature vectors consisting of P MFCCs to train models λ_i and $\bar{\lambda}_i$ associated with instrument i, and feature vectors consisting of $P' \neq P$ MFCCs to train models λ_j and $\bar{\lambda}_j$ associated with instrument j, and use the same decision rule as described above. Future work will consider such schemes which hold promise.

Finally, note that the recognition accuracy of different instruments should not be compared without prior normalization since inequivalent training/test sets were used.

4.3.2 SVM classification

The "one Vs one" approach was used to classify the five instruments. It was preferred to the "one vs all" approach for computational cost reasons, since the data to be considered to compute the optimal hyperplane separating one instrument from all others is much larger. Several kernels were tested, including linear, polynomial and Radial Basis Function. Polynomial parameters were also varied. Best results were achieved using linear and polynomial kernels. The used polynomial kernel has the form

% correct	MAP	one Vs all
Alto Sax	55.7	61.4
Bassoon	84.3	86.8
Bb Clarinet	30.2	32.8
Flute	67.7	68.0
Oboe	70.9	70.3

Table 3: Performance of isolated note classification with GMM.

$$K(\mathbf{x},\mathbf{y}) = (s\,\mathbf{x}.\mathbf{y} + c)^d. \quad (4)$$

Parameters s and c were chosen to be equal to 1 after testing. Parameter d was varied from 2 to 4. Recognition accuracy with the best tested kernels is shown in table 4 in terms of percentage correct. It seems that using the linear kernel is very advantageous, since it is computationally inexpensive and performs well in most cases.

% correct	Linear	Poly (d=2)	Poly (d=3)	Poly (d=4)
Alto Sax	73.4	69.2	69.9	69.0
Bassoon	88.0	88.0	87.2	87.6
Bb Clarinet	31.2	33.0	27.0	28.5
Flute	82.8	76.3	86.8	86.4
Oboe	66.9	66.4	74.8	75.9

Table 4: Performance of isolated note classification with SVM using linear and polynomial kernels.

4.4 Recognition on solo phrases

In this section, we present a study on the recognition of instruments playing solo phrases. It is shown that transforming the feature data with PCA enhances the classification performance. This is particularly important when dealing with data from commercial CDs as recording conditions may vary significantly.

4.4.1 The use of Principal Component Analysis (PCA)

PCA is often used in classification applications in order to reduce the dimensionality of the feature space [13]. In fact, it may be used to "de-noise" the signal in the sense that the most relevant information is concentrated in the first few components of the transformed feature vectors which correspond to directions of maximum energy. PCA was performed as follows. A subset of each instrument training data was taken to form a global training set where each instrument had the same number of feature vectors. The covariance matrix of this selected training data was computed and its Singular Value Decomposition (SVD) was processed yielding

$$\mathbf{R}_\mathbf{x} = \mathbf{U}\mathbf{D}\mathbf{V}^t,$$

where $\mathbf{R}_\mathbf{x}$ is the covariance matrix of the selected feature vectors from all instruments, $\mathbf{U}$ 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 $\mathbf{Y} = \mathbf{W}\mathbf{X}$, where $\mathbf{X}$ is the matrix whose columns represent the training feature vectors such that $\mathbf{X} = [\mathbf{x}_1,...,\mathbf{x}_\theta]$. The same transform matrix $\mathbf{W}$ was applied on test feature vectors.

4.4.2 Results and discussion

Recognition accuracy obtained with a 16-component GMM with and without PCA is shown in the first two columns of table 5. The overall recognition accuracy is improved (with up to 7.2 percentage point improvement for the Flute). Note that rescaling the data in order to get unit covariance in the transformed PCA space results

in very poor performance with GMM classification. Alternatively, Independent Component Analysis (ICA) was used in other work [7] to transform the feature space, yet we found no objective justification for its utilization. In fact, the improvement achieved thanks to ICA may only be due to the PCA pre-processing stage of the ICA algorithm.

Increasing the number of Gaussian components from 16 to 32 resulted in better accuracy except for the Alto Sax indicating that even more data should be used for training this instrument model.

The best overall recognition accuracy (obtained with PCA and GMM32) is 67.2 % which can be regarded as "satisfactory" considering our working constraints : not only a basic and limited set of features, namely MFCCs, was used, but also important variability was introduced in instrument instances, performers and recording conditions.

One way to improve the recognition accuracy is to increase the number of frames over which classification decisions are combined as it is suggested in table 6. In fact, when varying the decision length from 0.5 s to 10 s, significant enhancement is achieved (up to 21.9 percentage point improvement for the Alto Sax).

In our simulations, the GMM performed much better than SVM. Marques [9] reported better results with SVM but only two component densities were used in the Gaussian mixture as insufficient data was available for training. Thus, it turns out that GMM classification provides better recognition accuracy when important training sets can be used.

	GMM16	GMM16	GMM32
PCA		×	×
Alto Sax	62.1	63.2	55.5
Bassoon	43.8	50.3	53.7
Clarinet	50.0	46.7	52.0
Flute	62.1	69.3	78.6
Oboe	96.0	95.5	96.0

Table 5: Performance of solo phrases classification with GMM and PCA transformation (in % correct).

Test Length	1 s	5 s	10 s
Alto Sax	67.0	77.8	88.9
Bassoon	47.7	38.9	44.4
Clarinet	48.8	55.1	55.6
Flute	73.5	88.2	93.7
Oboe	96.3	98.3	98.4

Table 6: Performance of solo phrases classification with GMM16 and PCA transformation for different decision lengths (in % correct).

5. CONCLUSION

In this paper, recognition of musical instruments playing solo phrases was addressed. Two classification schemes, parametric and non-parametric, were considered, namely GMM and SVM, with a number of variations. Classifiers were trained on important data sets which allowed better performance with GMM. The use of PCA to transform the feature data was discussed and resulted in increased performance. Very high recognition accuracy, with larger test data, was achieved for the Flute and the Oboe under changing recording conditions and with several different instrument instances and performers.

Future work will consider more features better adapted to our task, which is a key issue for better overall performance. Additionally, more instrument classes will be considered and dynamic models such as Hidden Markov Models in association with SVM will be aimed at.

6. ACKNOWLEDGEMENTS

The authors wish to thank Olivier Cappé, Jean-François Cardoso and Claire Waast-Richard for fruitful discussions on current classification and speech recognition technologies.

REFERENCES

[1] Ircam studio online. http://www.ircam.fr.

[2] The university of iowa electronic music studios. http://theremin.music.uiowa.edu.

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

[4] 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.

[5] 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.

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

[7] Antti Eronen. Musical instrument recognition using ica-based transform of features and discriminatively trained hmms. In *Seventh International Symposium on Signal Processing and Its Applications*, pages 133–136, July 2003.

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

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

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

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

[12] F. Opolko and J. Wapnick. Mcgill university master samples. McGill University, 1987.

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

[14] D. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17:91–108, 19905.