# HIERARCHICAL CLASSIFICATION OF MUSICAL INSTRUMENTS ON SOLO RECORDINGS

*Slim ESSID, Gaël RICHARD and Bertrand DAVID*

GET - Télécom Paris (ENST)
37, rue Dareau - 75014 Paris - FRANCE

## ABSTRACT

We propose a study on the use of hierarchical taxonomies for musical instrument recognition on solo recordings. Both a natural taxonomy (inspired by instrument families) and a taxonomy inferred automatically by means of hierarchical clustering are examined. They are used to build a hierarchical classification scheme based on Support Vector Machine classifiers and an efficient selection of features from a wide set of candidate descriptors. The classification results found with each taxonomy are compared and analysed. The automatic taxonomy is found to perform slightly better than the "natural" one. However, our analysis of the confusion matrices related to these taxonomies suggest that both are limited. In fact, it shows that it could be more advantageous to utilise taxonomies such that the instruments which are commonly confused are put in distinct decision nodes.

## 1. INTRODUCTION

Various audio classification tasks have been addressed using hierarchical classification schemes. This approach is mainly motivated by the fact that more successful classification is thereby expected to be achieved, compared to "flat" systems. This is particularly true for the task of musical instrument classification. In fact, a number of previous studies have shown that it is advantageous to use a hierarchical framework for machine recognition of musical instruments, both on solo isolated notes [1, 2, 3, 4] and multi-instrument music [5].

In the solo music case, most works exploited intuitive taxonomies, roughly following the instrument families organisation, which have been exclusively tested on isolated notes. A very few attempts have been made at acquiring automatic taxonomies [4, 6], meant to be optimal with respect to classification performance. However, no study has compared the efficiency of automatic taxonomies against natural ones for the task of musical instrument recognition on solo performance recordings, especially in association with advanced machine learning techniques.

Such a comparison is proposed in this paper. We start by an overview of our hierarchical classification system. We then describe the taxonomies considered, and particularly how the automatic taxonomy is inferred. Subsequently, we compare the recognition rates and confusions made by the classification schemes based on these two alternatives. Finally, we suggest some conclusions.

Corresponding author's e-mail: slim.essid@enst.fr

## 2. OVERVIEW OF THE HIERARCHICAL CLASSIFICATION SYSTEM

### 2.1. Feature extraction and selection

A wide selection of more than 540 signal processing features is considered. Since these features and their extraction have been extensively described in various previous work in the field of Music Information Retrieval (see [7] for example), in the following, we merely list the attributes which were examined in our study.

*Temporal features* (95 features) consist of autocorrelation coefficients, features obtained from the statistical moments, zero crossing rates, and amplitude modulation features.

*Cepstral features* (183 features) are mel-frequency cepstral coefficients, and cepstral coefficients obtained from a constant $Q$ transform [8], as well as their first and second time derivatives.

*Spectral features* (101 features) include features obtained from the statistical moments, MPEG-7 audio spectrum flatness, spectral irregularity, spectral crest, spectral slope, spectral decrease, frequency cutoff, temporal variation of spectrum, and octave band signal intensities and their ratios providing a coarse description of the energy distribution of sound partials [9].

*Wavelet features* (35 features) obtained from the statistics of wavelet coefficients.

*Perceptual features* (129 features) are also utilised, namely loudness, sharpness, spread and signal to mask ratios.

In order to fetch the most useful features among all the candidates considered, automatic feature selection is used [10]. We have developed a new algorithm which have proved to be efficient on audio data, compared to the state-of-the-art methods. The first step of our algorithm is to cluster all features considered (from all classes) so that the most redundant ones are put in the same clusters. Then, in each cluster the most valuable feature is selected using the weights estimated, for each attribute, by a Linear Discriminant Analysis (LDA) [11] (*i.e.* the feature which has the first rank is selected). The obtained features are thus the most useful representatives of the feature clusters.

The motivation for this approach is that LDA-based feature selection is very successful at producing features ensuring high class separability but makes no account for another important requirement, *i.e.* selecting non-redundant features. Therefore, by picking features from different clusters, we obtain both highly discriminative and non-redundant attributes. Furthermore, an automatic taxonomy of features is obtained by this method, which is very interesting from an analysis point of view. More details on this algorithm will be given in future work.

### 2.2. Classification

Given a $L$-level hierarchical taxonomy (such as the one depicted in Figure 1, where $L$=3), hierarchical classification can be seen as a procedure combining $L$ flat classification schemes, each related to one level of the taxonomy. At each level $l$, unknown sounds are assigned, by the corresponding flat classification scheme, to a particular class $\hat{\Omega}_l$ which is one of the nodes of the taxonomy. The final decision is reached by following the path, from the root node (N0) to the leafs (at the bottom), found by crossing at each level $l$, the selected node $\hat{\Omega}_l$. When the bottom of the taxonomy is reached, the sounds are classified among the possible instrument classes.

At each level, we use Support Vector Machine classifiers (SVM) [12] with a Gaussian kernel, based on the features selected for the discrimination of the original classes. SVM are used in a "one vs one" fashion, with probabilistic outputs [13] and the Hastie & Tibshirani technique for coupling the pairwise decisions [14]. This allows for using the usual Maximum A Posteriori (MAP) decision rule [11].

## 3. TAXONOMIES OF MUSICAL INSTRUMENTS

### 3.1. Instrument families

Various taxonomies have been proposed for musical instrument classification on isolated notes roughly following the instrument families organisation [1, 2, 3]. While some declinations are common to these studies, as for example the primary division of instruments into "sustained" and "pizzicati", other groupings are not unanimously shared, especially for the wind instruments.

We select a taxonomy inspired by [3] which organises the instruments with respect to the sound production mode. The restriction of this taxonomy to the instruments considered in this study (presented in Table 2) is depicted in Figure 1.

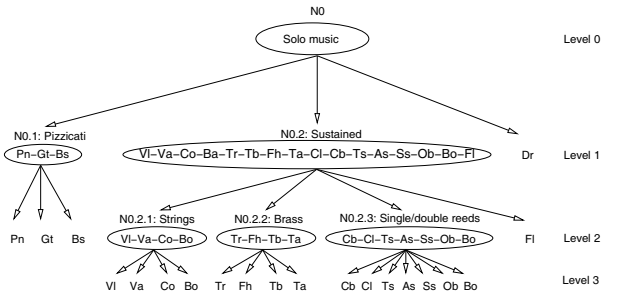


**Fig. 1**. A taxonomy following instrument families organisation. The related classification scheme will be referred to as FHC.

### 3.2. Inferring an automatic taxonomy

We now briefly describe our algorithm for inferring hierarchical taxonomies, originally presented in [6] (see [5] for a more detailed description).

The target instrument classes are organised using a hierarchical clustering algorithm. We use agglomerative hierarchical clustering [11] to produce "a hierarchy of nested clusterings". This is known to be an optimal and natural way of arranging the data in the sense that the most similar classes with respect to the chosen closeness criterion are then put in the same clusters.

The choice of the closeness criterion is critical. We need robust distances enabling us to reduce the effect of feature noise on the clustering performance. Also, the distances are required to be matched with the behaviour of the classifiers to be used. A convenient and robust means of measuring the closeness or separability of data classes is to use probabilistic distance measures between them, *i.e* distances between their probability distributions [11]. This is an interesting alternative to classic Euclidean distance between feature vectors known to be inefficient for sound source classification. For improved robustness, a kernel method is used to measure the probabilistic distances in a transformed features space (using a Gaussian kernel). We follow Zhou & Chellapa's approach to measure two alternative distances: the divergence and Bhattacharryya distances [15]. The divergence turns out to entail the best clustering performance.

## 4. EXPERIMENTS

The instruments used in our study are given in Table 2. Solo (unaccompanied) music was excerpted from commercial recordings of each instrument. The properties of the sound database assembled are summed up in Table 1. Note that there is a complete separation between sources[1] from which the training excerpts were extracted and those providing the testing excerpts. This allows for assessing the generalisation capabilities of the classification schemes. We were unfortunately unable to assemble enough data for the bass clarinet. Consequently, all the excerpts which were available for this class were used for training, hence the bass clarinet is not tested but it is considered as a possible class when testing the other instruments.

| Instrument | Train src. | Train | Test src. | Test |
|---|---|---|---|---|
| Pn | 7 | 22' 16" | 7 | 14' 13" |
| Gt | 5 | 10' 43" | 5 | 15' 58" |
| Bs | 3 | 7' 37" | 5 | 12' 44" |
| Ba | 3 | 6' 44" | 4 | 6' 45" |
| Co | 5 | 15' 47" | 5 | 12' 7" |
| Va | 5 | 16' 37" | 5 | 15' 57" |
| Vl | 6 | 34' 11" | 5 | 24' 11" |
| Ta | 2 | 2' 49" | 2 | 1' 51" |
| Tb | 4 | 15' 28" | 4 | 7' 1" |
| Fh | 4 | 3' 43" | 2 | 3' 24" |
| Tr | 5 | 10' 46" | 5 | 11' 30" |
| Bo | 4 | 13' 0" | 4 | 12' 14" |
| Ts | 3 | 11' 13" | 5 | 6' 40" |
| As | 3 | 20' 7" | 4 | 10' 15" |
| Ss | 2 | 13' 49" | 2 | 7' 51" |
| Fl | 5 | 16' 31" | 5 | 15' 56" |
| Ob | 4 | 14' 46" | 5 | 14' 40" |
| Cl | 5 | 8' 34" | 5 | 13' 38" |
| Cb | 4 | 2' 13" | 0 | 0' 0" |
| Dr | 3 | 3' 1" | 1 | 4' 24" |

**Table 1**. Sound database used. "Train src." and "Test src." are respectively the number of different sources used, "Train" and "Test" are respectively the total lengths (in minutes and seconds) of the train and test sets. Train data size is sometimes smaller than the test data size, since this database is used in a larger study where it is necessary to keep a part of the training data as a hold-out set (not appearing here) for tuning purposes.

[1] a *source* is a music recording such that, either the recording studio, the artist or the instrument instance differs from one source to another.

40 features from the 543 candidates were selected using our feature selection algorithm. Spectral features are largely represented in the subset of the most valuable features (18/40 are spectral features). Among the other successful features are 9 cepstral attributes and 6 perceptual ones.

| Instrument | Code | Instrument | Code |
|---|---|---|---|
| Alto sax | As | Oboe | Ob |
| Tenor sax | Ts | Soprano sax | Ss |
| Bassoon | Bo | Piano | Pn |
| Double bass, *pizzicato* | Bs | Double bass *arco* | Ba |
| Bass clarinet | Cb | Tuba | Ta |
| Clarinet | Cl | Trombone | Tb |
| Cello | Co | Trumpet | Tr |
| Flute | Fl | Viola | Va |
| French horn | Fh | Violin | Vl |
| Guitar | Gt | Drums | Dr |

**Table 2**. Instruments considered and their codes.

The classification system based on the taxonomy following instrument families will be referred to as FHC, while the one based on the automatic taxonomy will be referred to as AHC.

Recognition success is evaluated over a number of decision windows. Each decision window combines elementary decisions taken over $N_t$ consecutive 32-ms analysis frames. The recognition success rate, for each class, is the percentage of successful decisions over the total number of available decision windows. In our experiment we use $N_t = 249$, which corresponds to 4-second decisions (there is 50% overlap between analysis windows). Note that these are the final decisions, *i.e.* the decisions made at the leaf nodes. In fact, the decisions at intermediate nodes are taken over a single analysis frame, hence the rates which will be presented for these intermediate decisions are measured with $N_t = 1$.

### 4.1. A reference "flat" classification system

A flat classification system based on Gaussian Mixture Models (GMM) [11] was built to serve as a reference. We use these classifiers instead of the "one vs one" classification scheme based on SVM, which is used in the hierarchical systems studied, to reduce the computational load. In fact, 190 pairwise classifiers, targeting all possible pairs of instrument classes among 20, would be required. A GMM with 8 component densities was thus computed for each of the 20 instrument classes considered based on the 40 features selected. The average accuracy found is 61.3%. In addition to the "usual" intra-family confusions, others between instruments from distinct families are made by the system. For example, the bassoon is confused with the French horn 24.3% of the time, the oboe is confused with the trumpet 11.2% and the tenor sax assigned to the class violin 24.9% of the time.

### 4.2. The automatic taxonomy obtained

The taxonomy obtained based on the 40 features selected is depicted in Figure 2. Note that it is different from the one acquired in [6] as a wider selection of features is used, which allows for describing further qualities of the instrument sounds. The groupings do not always reflect the instrument families organisation, but they are consistent with the confusions which are found by the flat instrument classification system, in the sense that the instruments which are the most frequently confused are found in the same nodes of the taxonomy. For example, the trumpet and the oboe are grouped together in the node N0.3, and the French horn and the bassoon are both in the nodes N0.2 and N0.2.1.

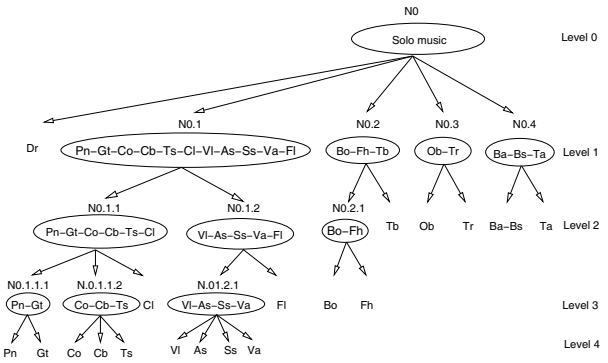


**Fig. 2**. Automatic taxonomy of musical instruments. The related classification scheme is referred to as AHC.

### 4.3. Comparing the performance of the two taxonomies

Table 3 sums up the recognition accuracies achieved by the 2 hierarchical classification systems.

Both of the hierarchical systems perform[2] better than the flat system. On average, the automatic taxonomy yields better results than the instrument families taxonomy. AHC identifies more successfully 11 instruments among 18, compared to FHC. However, the difference in average accuracy is quite small. Let us first analyse the confusions at the intermediate nodes.

**With FHC**. The average accuracy at the top level (level 1) is 82.3%. The instruments of the node N0.2 are successfully recognised 96.4% of the time. The guitar is assigned to N0.2 30% of the time and so are the drums in 22.8% of the cases. The classification among the nodes of the second level is more difficult. The average accuracy is 84.0% at the node N0.1 and only 64.3% at N0.2. The most important confusions occur for the flute and the tuba which are assigned to N0.2.3, respectively 43.4% and 41.8% of the time. Finally, at the last level (level 3), important confusions occur for the viola with the violin, the clarinet with the alto sax, the soprano sax with the alto sax, and the French horn with the trombone, more than 30% of time.

**With AHC**. The average recognition accuracy at the top level (level 1) is 74.4%. Instruments of the node N0.2 are successfully classified only 56.2% of the time, they are frequently assigned to N0.1. While with FHC many confusions occur at the second level, the average recognition accuracy is over 80% at the second level of the automatic taxonomy. The main difficulty arise for the classification of the clarinet which is assigned to N0.1.2 48.2% of the time. In fact, most confusions occur between instruments which belong to the same instrument family but were put by the clustering algorithm in different groups. At the last level, one can point out the confusions of the alto sax with the soprano sax (66.1% of the time) and the bassoon with the French horn (56.0% of the time).

By studying the confusion matrices output by the two systems (not given here for the lack of space), one can observe that the use

[2]Performance is measured in terms of the average recognition accuracy.

of a particular taxonomy results in a different distribution of the instrument confusions, without necessarily achieving a substantial improvement in overall performance (as required). Surprisingly, these confusions do not reflect the expectation that instruments grouped together in the same nodes of the taxonomy are less frequently confused once the leaf nodes are reached. It seems more advantageous to have instruments which are difficult to discriminate in different decision nodes at the bottom of the hierarchy. For example, the flute is confused with the oboe 11.3% by FHC but only 4.1% of the time by AHC, as the flute and the oboe are found in distant nodes of the automatic taxonomy. This is also true for the pair bassoon vs French horn: these two instruments form the node N0.2 of the automatic taxonomy while they are in different nodes of the instrument family taxonomy (N0.2.2 and N0.2.3).

| % correct | Families | Automatic |
|---|---|---|
| Pn | 93.9 | **95.2** |
| Gt | 74.5 | **77.3** |
| Ba-Bs | **93.7** | 93.5 |
| Co | 58.0 | **59.0** |
| Va | 61.0 | **61.6** |
| Vl | 66.6 | **70.2** |
| Ta | 34.6 | **37.9** |
| Tb | **69.5** | 67.7 |
| Fh | 55.3 | **64.4** |
| Tr | 71.0 | **74.1** |
| Bo | **57.9** | 43.6 |
| Ts | **18.7** | 18.1 |
| As | **95.0** | 93.9 |
| Ss | 8.4 | **9.4** |
| Fl | 65.5 | **77.9** |
| Ob | **91.3** | 88.2 |
| Cl | **42.6** | 39.4 |
| Dr | **91.2** | 90.7 |
| Mean | 63.8 | **64.6** |

**Table 3**. Instrument recognition accuracies fond by the classification schemes tested. "Families" and "Automatic" refer to the hierarchical systems based respectively on the instrument families and automatic taxonomies.

## 5. CONCLUSIONS

In this paper, we have analysed and compared the performance of two alternative hierarchical taxonomies for the task of instrument recognition on solo recordings: the first is inspired by the "natural" instrument families organisation and the second was built automatically by hierarchical clustering, grouping together the instrument having similar acoustic features into the same clusters. Using SVM classifiers we found that the automatic taxonomy performed only slightly better than the natural one.

Analysing the confusion matrices related to each of the classification schemes, we arrived at the surprising conclusion that when instruments which are difficult to discriminate are grouped by the taxonomy in the same **decision nodes** at early levels, they are not more accurately classified by either classification scheme. This finding contradicts the fact that "close" classes should be systematically grouped together when building a taxonomy for classification. It implies that it might be useful to consider acquiring taxonomies which spread the instruments which are difficult to discriminate over distant nodes. While this could be a limitation if considering generative classification approaches (such as GMM), since it is difficult to model the probability densities of the resulting class-groups (potentially heterogeneous), it would not be necessarily a shortcoming if non-linear SVM classifiers are used, as complex decision surfaces can be learned.

## 6. REFERENCES


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

[2] Antti Eronen, "Automatic musical instrument recognition," M.S. thesis, Tampere University of Technology, April 2001.

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

[4] T. Kitahara, M. Goto, and H.G. Okuno, "Category-level identification of non-registered musical instrument sounds," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada., May 2004.

[5] Slim Essid, Gaël Richard, and Bertrand David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Transactions on Speech and Audio Processing*, January 2006, to appear.

[6] Slim Essid, Gaël Richard, and Bertrand David, "Inferring efficient hierarchical taxonomies for MIR tasks: Application to musical instruments," in *6th International Conference on Music Information Retrieval (ISMIR)*, London, UK, September 2005.

[7] Geoffroy Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Tech. Rep., IRCAM, 2004.

[8] Judith C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *Journal of the Acoustical Society of America*, vol. 105, pp. 1933–1941, March 1999.

[9] Slim Essid, Gaël Richard, and Bertrand David, "Musical instrument recognition based on class pairwise feature selection," in *5th International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

[10] I. Guyon and A Elisseeff, "An introduction to feature and variable selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

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

[12] Christopher J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Journal of Data Mining and knowledge Discovery*, vol. 2, no. 2, pp. 1–43, 1998.

[13] John C. Platt, "Probabilistic outputs for support vector machines and comparisions to regularized likelihood methods," *Advances in Large Margin Classifiers*, 1999.

[14] Trevor Hastie and Robert Tibshirani, "Classification by pairwise coupling," in *Advances in Neural Information Processing Systems*. 1998, vol. 10, The MIT Press.

[15] S. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: probabilistic distance measures in reproducing kernel hilbert space," *IEEE transactions on pattern analysis and machine intelligence*, to be published.