INSTRUMENT RECOGNITION IN POLYPHONIC MUSIC

Slim ESSID, Gaël RICHARD and Bertrand DAVID

GET - ENST (Télécom Paris) - TSI
46, rue Barrault - 75634 Paris Cedex 13 - FRANCE

ABSTRACT

We propose a method for the recognition of musical instruments in polyphonic music excerpted from commercial recordings. By exploiting some cues on the common structures of musical ensembles, we show that it is possible to recognize up to 4 instruments playing concurrently. The system associates a hierarchical classification tree with a class-pairwise feature selection technique and Gaussian Mixture Models to discriminate possible combinations of instruments. Successful identification is achieved over short-time windows, enabling the system to be employed for segmentation purposes.

1. INTRODUCTION

The scope of the present work is targeted on machine recognition of musical instruments in a polyphonic context. This issue has been hardly addressed as most studies were dedicated to the recognition of musical instruments playing isolated notes [1, 2]. Fewer contributions were concerned with the recognition on solo musical phrases involving a single instrument executing any musical score [3, 4]. Few attempts were made on a polyphonic content where many instruments play simultaneously. Given the complexity of this problem, a number of restrictions were considered in such studies either regarding the number of instruments to be recognized, the nature of music, or the musical notes played. In fact, recognition was often related to a source separation task requiring the knowledge or estimation of the pitches of the different notes [5, 6, 7]. Then, limitations arise due to the difficulty to extract the fundamental frequencies in the multi-pitch case, especially for octave-related notes. Moreover, such studies processed artificially mixed musical elements such as notes, chords or melodies. Using realistic musical recordings, Eggink & Brown proposed a system based on a missing feature approach [8] capable of identifying 2 instruments playing simultaneously. More recently, the same authors presented a system recognizing a solo instrument in the presence of musical accompaniment by recurring to the extraction of the most prominent fundamental frequencies in the audio signals [9].

In this paper, we introduce a multi-instrument recognition scheme processing real-world music, based on *a priori* knowledge of the musical context. Our proposal does not require any pitch-detection or separation operations. We show that it is possible to recognize up to 4 instruments playing concurrently on the basis of realistic musical hypotheses. We choose a piano jazz quartet ensemble to illustrate the performance of the proposed methodology.

Our approach follows a hierarchical classification scheme where a number of classes to be recognized can be grouped together at high levels of the suggested taxonomy. These classes consist of particular combinations of instruments that are deduced from the musical genre. We show through experimental work that high recognition accuracy can be achieved with up to 4 instruments playing at the same time.

The outline of the paper is the following. We first present the recognition system architecture. Subsequently, we give a description of the signal processing features used and our algorithm for selecting the most relevant features. We then briefly describe the one vs one GMM classifier employed. Finally, we proceed to the experimental validation and suggest some conclusions.

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

2. SYSTEM ARCHITECTURE

2.1. Music instrumentation cues

Choosing a specific music instrumentation for a composition is one of the degrees of freedom of a composer. If in contemporary music (especially in classical and jazz) a large variety of instrumentations are used, it is clear that most trio and quartet compositions use typical instrumentations traditionally related to some musical genre. For example, typical jazz trios are composed of piano or guitar, double bass and drums and typical quartets involve piano or guitar, double bass, drums and a wind instrument or a singer. In a vast majority of musical genres, each instrument, or group of instruments, has a typical role related to rhythm, harmony or melody. Not surprisingly, a number of studies in Audio Scene Analysis of modern music aim at extracting sub-symbolic representations such as the hierarchical beat structure and the drum patterns (related to the rhythmic part), the chord changes and bass line (related to harmony) and the melody (see [10] for example). When jazz/pop ensembles of 4 or less instruments are considered, namely solos to quartets, rhythm is often carried out by percussive instruments (drums or percussions) and/or a bass instrument (such as double bass) while melody or second line is often played by a monophonic instrument (sax, trumpet, voice, etc.). Polyphonic instruments such as the piano or the guitar have more diverse roles since they can be involved in the harmony, rhythm or melody parts.

2.2. The taxonomy

Our approach consists in defining classes inspired by the possible instrumentations related to the musical genre. These classes are instruments or groups of instruments possibly playing simultaneously at certain parts of a musical piece. The number of possible combinations is reduced by building super-classes consisting of unions of classes having similar acoustic features. They constitute the top-level of the hierarchical classification scheme depicted in figure 1. The scheme is given for the jazz example. Nevertheless,

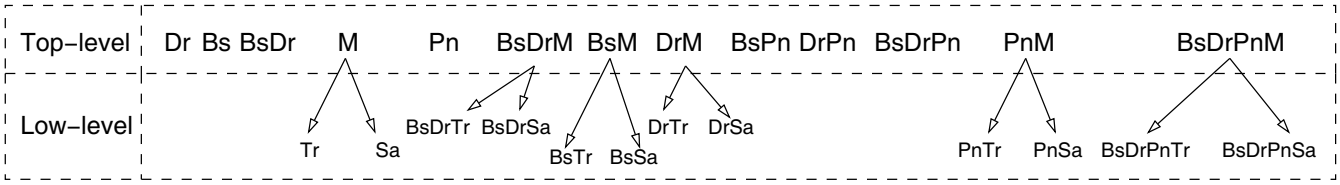

Fig. 1. An example of taxonomy for the recognition of musical instruments in jazz piano ensembles from solos to quartets. *Dr :drums, Bs :double bass, M : Monopitched instrument, Pn :piano*.

it is believed that the suggested methodology is not restricted to the presented example. The only constraint is that the genre of the music to be processed is known. Such an indication can be obtained in a first stage either by exploiting the textual metada accompanying the audio or by using a musical genre recognition system. In fact, once the genre is known, one can easily adapt the alternative instrument combinations corresponding to each super-class by exploiting information on the possible instrumentations found in real music.

The classification is performed hierarchically in the sense that a given test segment is first classified among the top-level classes, then it is determined more precisely (when needed) in the lower-level. For example, if a test segment involves double bass, drums and trumpet, then it is first identified as **BsDrM** and subsequently as BsDrTr, where Bs stands for double bass, Dr for drums, M for a wind instrument (sax or trumpet) and Tr is Trumpet.

In the following, we describe the aspects related to the signal processing features and classifiers used within this architecture.

3. ON FEATURES AND THEIR SELECTION

There exists no widely shared consensus about an optimal feature set dedicated to the instrument recognition task. One of the main difficulties is then to choose among various descriptors proposed in the literature [1, 2, 4]. For this purpose, the so-called IRMFSP (Inertia Ratio Maximization with Feature Space Projection) is exploited in order to fetch the most relevant features from a very wide initial set of potentially useful ones. The particularity of our approach is that IRMFSP is performed in a class-pairwise manner [4].

3.1. The whole feature set

The classical features considered in this work are briefly described hereafter, keeping in mind that they were chosen for the robustness of their extraction from polyphonic music recordings.

The autocorrelation coefficients, temporal statistics obtained from the first 4 statistical moments and the Zero Crossing Rates characterize the waveform in the time domain.

The MFCC (Mel-Frequency Cepstral Coefficients) tend to represent the spectral envelope and its variations over successive frames (first and second derivatives also used).

Spectral Centroïd, Spectral Width, Spectral Asymmetry and Spectral Flatness [11] and their time derivatives describe the spectral statistic distribution and its evolution over time. The MPEG-7 Audio Spectrum Flatness [12] and the derivatives of the constant-Q coefficients are added. Finally, the frequency below which 99% of the spectral energy is accounted, the Frequency cutoff, is computed.

Since pitch estimation has to be avoided, a new feature set is derived to roughly evaluate the power distribution among the different harmonics [4]. The log energy of each subband of an octave triangular filter bank is computed leading to the OBSI (Octave Band Signal Intensities) vector. The vector OBSIR (OBSI Ratios) is then obtained by forming the quotient of each subband intensity to its previous.

3.2. Feature selection

The feature selection is performed class-pairwise, leading to an optimized subset of features for each possible pair of classes. The subsequent classification is processed in a one vs one scheme. This approach has proven to be more successful than the classic one where a single set of attributes is used for all classes [4].

Several techniques are available to perform the selection task as the family of Sequential Feature Search Techniques or Genetic Algorithms [13]. In this work, we retained an intuitive approach, the IRMFSP, which has been found successful in a musical processing context.

The IRMFSP 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) of the Between-class inertia B_i to the "average radius" of the scatter of all classes R_i defined as :

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

The idea behind is to select features that enable good separation between classes with respect to the within-class spreads. The selected additional feature corresponds to the highest ratio r_i. Using this criterion may result in redundant feature subsets wherein the same signal attributes are embedded in a number of features still entailing high r_i-values. Then the algorithm is modified as in [14] to take into account the non-redundancy constraint by introducing an orthogonalization step at each feature selection iteration. In summary, at each iteration,

- the ratio r_i is maximized yielding a new feature subset $\mathbf{X}_i$,

- the feature space spanned by all observations is made orthogonal to $\mathbf{X}_i$.

The algorithm stops when the ratio r_d measured at iteration d gets much smaller than r_1, *i.e.* when $\frac{r_d}{r_1} < \epsilon$ for a chosen ϵ, which means that the gain brought by the last selected feature has become non-significant. This provides a convenient means for implicitly selecting the number of useful features when the size of the feature subset to be selected is not a constraint.
The process of selecting the best features for discriminating between any possible pair of classes using IRMFSP will be referred to as C_2^K-IRMFSP. Note that there are as many feature subsets selected as class pairs and their sizes are not necessarily equal. Whenever two classes can be easily distinguished, the number of needed features is expected to be smaller.

4. ON CLASSIFICATION

The classification is performed through a one vs one scheme using Gaussian Mixture Models. As the GMM has been extensively used for various classification tasks since their application to text-independent speaker recognition (see [15] for example), only its pairwise utilization is here discussed.
As many GMM classifiers as instrument pairs are built based on different feature subsets found thanks to C_2^K-IRMFSP. 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 at time t if

$$p(\mathbf{x}_t|\Omega_i) > p(\mathbf{x}_t|\Omega_j) \quad (2)$$

where $\mathbf{x}_t$ is the test feature vector observed at time t and $(p(\mathbf{x}_t|\Omega_k))_{k=i,j}$ is the class-conditional probability of $\mathbf{x}_t$, which is modeled as a GMM.

5. EXPERIMENTAL STUDY

5.1. The sound database

We collected sound excerpts corresponding to the classes given in figure 2 from both commercial Compact Disc (CD) recordings and RWC jazz music database [16]. It was found that, in practice, some combinations appearing on figure 1 are very rare. Consequently, they were not tested in the experimental study, since no sound material representing these classes could be assembled. Table 1 gives an overview of the data used in our experiments. 2/3 of the data was used for training and the remaining 1/3 used for testing.

Fig. 2. Experimented taxonomy for the recognition of musical instruments in jazz piano ensembles from solos to quartets.

5.2. The selected features

A total of 164 features are explored for our classification task. Class-pairwise feature selection is performed at each level of the taxonomy. At the top-level (the parents level consisting of 8 classes), the average number of features selected using C_2^8-IRMFSP (with $\epsilon = 10^{-6}$) is 30.33. The size of the most relevant feature sets ranges from 6 for the pairs {**M**, BsPn} to 52 for the {Dr, BsDr} confrontation. The most successful features are MFCC, OBSI, spectral statistics, Zero Crossing Rates and frequency cut-off. At the low-level, 37 features are selected for discriminating between the sax and the trumpet, 25 for the pair BsDrSa/BsDrTr and 44 for BsDrPnSa vs BsDrPnTr.

	Training (s)	Test (s)
Pn	777	388
Sa	555	278
Tr	1471	738
Dr	308	154
BsPn	356	178
BsDr	184	92
BsDrSa	173	86
BsDrTr	114	57
BsDrPn	1142	571
BsDrPnSa	98	49
BsDrPnTr	518	259

Table 1. Sound database for the experimental study. 'Training' and 'Test' are respectively the total durations of training and test material in seconds.

5.3. Classification results

Table 2 presents the recognition accuracies at the 2 levels of the taxonomy, obtained with one vs one GMM classifiers (with 64 component densities) using the features found previously. The rates presented between parenthesis are the ones corresponding to the accuracies found within the same level of the hierarchy, independently from the recognition success of the parent nodes. Scoring is performed as follows : for each test signal, a decision regarding the instrument classification is taken every T seconds (L observations). The recognition success rate is then, for each class, the percentage of successful decisions over the total number of T-second test segments. Two decision-lengths are tested, $T = 2s$ and $T = 4s$.

Very satisfactory results are found for all classes except the drums and BsDrPn. With $T = 2s$, the drums are confused with BsDrPn 68.66% of time and with the class **M** in 13.43% of the cases. It is believed that the identification of this class could be highly improved by means of appropriate attributes such as wavelet descriptions. The class BsDrPn is classified as **BsDrPnM** 46.42% of the time. This is probably due to the fact that an important number of audio segments from the class BsDrPn may slip into the training set relative to **BsDrPnM**. Indeed, when assembling the experimental data, one is obliged to segment by hand **BsDrPnM** music in order to drop all material not involving the four instruments simultaneously, which is not so obvious given the temporal resolution of one's ear that is greater than the frame-size of 32ms.

With longer decision-lengths, better performance is achieved in average. Some classes, such as BsDr and BsPn are always identified correctly. However, this parameter can be a sensitive one in segmentation applications. In fact, since short-time decisions can be taken ($\approx 2s$), the proposed system can be easily employed in a task of segmentation of music duos, trios and quartets. By combining the decisions given over 2s-windows, it is easy to define

% correct	$T \approx 2s$	$T \approx 4s$
Dr	10.45	12.12
BsDr	97.37	100.00
M	99.39	99.70
Pn	95.31	95.83
BsDrM	95.88	95.83
BsPn	97.65	100.00
BsDrPn	53.58	54.55
BsDrPnM	93.70	93.65
Sa	96.80 (97.39)	98.31 (98.61)
Tr	93.12 (93.69)	93.60 (93.88)
BsDrSa	95.88 (100.00)	95.83 (100.00)
BsDrTr	85.90 (89.66)	88.99 (92.86)
BsDrPnSa	86.70 (92.52)	90.12 (96.23)
BsDrPnTr	88.70 (94.74)	93.65 (100.00)

Table 2. Recognition accuracies with 2-s and 4-s decision lengths.

% correct	T ≈ 2s
1 source	97.58 (54.02)
2 sources	97.51
3 sources	74.73
4 sources	93.70

Table 3. Correct detections of the number of musical sources. For 1-source detection, the value between parenthesis is found when considering the drums as a possible source.

the segments wherein each instrument or group of instruments is involved.

Another application of this work is the detection of the number of instruments or sources involved in the audio signal, which can be very helpful for source separation tasks. Table 3 gives the percentage of correct detections of the number of sources using 2-s decision lengths. The average success rate is 90.88 %.

6. CONCLUSION

A new approach to machine recognition of musical instruments has been proposed addressing the polyphonic musical context. The suggested taxonomic scheme, which is inspired by intuitive cues on the structure of musical ensembles, is capable of identifying up to 4 instruments playing concurrently. Using class-pairwise selected subsets of features and GMM classifiers exploited in a one vs one fashion at each level of the hierarchy results in high recognition accuracies. The system does not need to perform any fundamental frequency analysis and can be used to detect the number of sources to help source separation tasks. Furthermore, classification can be done over short-time windows (2s), which allows us to perform segmentation.
Future work will be dedicated to the assessment of the proposed methodology on more varied musical genre and instrumentation.

7. REFERENCES

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

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

[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*, vol. 109, pp. 1064–1072, March 2000.

[4] Slim Essid, Ga el 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.

[5] Kunio Kashino and Hiroshi Mursae, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Communication*, vol. 27, pp. 337–349, September 1998.

[6] Tomoyoshi Kinoshita, S. Sakai, and Hidehiko Tanaka, "Musical sound source identification based on frequency component adaptation," in *IJCAI Workshop on Computational Auditory Scene Analysis (IJCAI-CASA)*, Stockholm, August 1999.

[7] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," vol. 92, no. 4, pp. 712–729, April 2004.

[8] Jana Eggink and Guy J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, April 2003, pp. 553–556.

[9] Jana Eggink and Guy J. Brown, "Instrument recognition in accompanied sonatas and concertos," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montréal, Canada, May 2004, pp. 217–220.

[10] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," in *Speech Communication*, corrected proof available on line August 2004 In press, Ed., 2004.

[11] Slim Essid, Ga el Richard, and Bertrand David, "Efficient musical instrument recognition on solo performance music using basic features," in *AES 25th International Conference*, London, UK, June 2004.

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

[13] F. J. Ferri, P. Pudil, Mohamad Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," *Pattern Recognition in Practice IV*, pp. 403–413, 1994.

[14] 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.

[15] Douglas A. Reynolds and Richard C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[16] Masataka Goto, Hiroki Hashigushi, Takuishi Nishimura, and Ryuichi Oka, "RWC music database: Popular, classical, and jazz music databases," in *International Conference on Music Information Retrieval (ISMIR)*, Paris, France, October 2002.