Recognition of Musical Instruments by Statistical Classification

GIOVANNI COSTANTINI, DANIELE CASALI
Dept. of Electronic Engineering
University of Rome “Tor Vergata”
Via del Politecnico 1, 00173 – Rome
ITALY

Abstract: - The correct classification of single musical sources is a relevant aspect for the source separation task and the automatic transcription of polyphonic music. In this paper, we present a classification experiment on six different musical instruments: violin, clarinet, flute, oboe, saxophone and piano. It is characterized by two steps. In the first step, a suitable signal preprocessing based on FFT and QFT (Q-constant Frequency Transform) is adopted for feature extraction and data set preparation. In the second step, a nonexclusive classification method is proposed to handle the inevitable overlapping among classes. It is obtained by a co-operative clustering technique. The success of this kind of classification method is conditioned by the adopted clustering procedure. We propose a hierarchical scale-based approach for this task, carrying out good results.

Key-Words: - Pattern Recognition in Music, Musical Instruments, Parameter Estimation, Spectral Analysis.

1 Introduction
In acoustic and musical context, source separation and recognition problem are very relevant to process single source sound independently of background and to automatically transcribe polyphonic music. Several attempts in this direction have been recently made [2,3,10,11].

Since musical instruments can be played in very different ways and sounds of different instruments can have similar characteristics, the classification of single sound sources by means of their characteristic parameters could be the first step to be made for the source separation task. Hence, the success of a recognition method requires the use of appropriate preprocessing algorithms, capable of extracting all the distinctive features of the musical signals, and the application of classification methods operating in a nonexclusive environment. The preprocessing method is described in Sect. 2. It constitutes a preliminary attempt to model musical signals.

A consequence of the fact that musical instruments can be played in various manners is that the patterns representing the musical signals in a feature space are contained in overlapping regions. The correspondence instrument/region is therefore confusing: i.e., there are not sharp boundaries among the regions associated with different instruments. Due to this fact, the use of a nonexclusive classifier [7] could be desirable. Such a kind of classification technique can be obtained by clustering each class independently. Namely, a K-class supervised problem can be reduced to K distinct unsupervised ones. The classification results strongly depend on the success of the adopted clustering procedure. In the present work, the clustering procedure is optimized by a hierarchical scale-based approach, as it will be explained in Sect. 3. In Sect. 4 experimental results will be reported and discussed.

2 Musical Signal Preprocessing
A frequency analysis is first performed on notes played by musical instruments, in order to detect the signal harmonics. Using the Fast Fourier Transform (FFT), which is very rapid and therefore suitable for real-time calculations, the frequency resolution could be not sufficient. In fact, a FFT with 4096 temporal samples x[n] on a sound recorded with the usual sampling rate (SR) of 44100 Hz, carries out a resolution of about 10.8 Hz between two FFT samples. This is not sufficient for low frequency notes, where the distance between two adjacent semitones is about 8 Hz (C3, 131 Hz and C#3, 139 Hz). The frequency resolution will get better if a higher number of temporal samples are used (with 8192 samples the resolution is of about 5.4 Hz), but that requires larger temporal windows for a fixed SR. In this case, the analysis of the instantaneous spectral information of the musical signal makes worse. To solve this problem, a Q-constant Frequency Transform (QFT) is used to detect the fundamental frequency of the note. Then, the upper harmonics may be individuated easily, as they are located at frequencies nearly multiples of the fundamental frequency.

In QFT transform, described in [1], the frequency channels are not linearly spaced – as in the FFT – but logarithmically spaced. It suits better for musical notes that are based on a logarithmic scale. Moreover, the QFT uses a varying number of temporal samples N[k] for calculating the frequency transform, more samples for calculating the transform of lower notes than of higher notes. This adaptive windowing yields to a good frequency resolution for all the notes of the tempered scale. As in a musical octave there are 12 semitones logarithmically spaced, the frequency value between two adjacent semitones increments of about 6 %. As an example, if we...
consider two notes C4 (261.6 Hz) and C#4 (277.2 Hz), we have an increase of \( \frac{277.2 - 261.6}{261.6} = 0.0596 \), which corresponds to a Q-value of \( Q = \frac{f}{\Delta f} = \frac{f}{(0.0596 - f)} \approx 17 \).

In the present work, a Q-value of 34 has been used to obtain 2 channels for every semitone. For the signal QFT calculation, it is necessary to establish two other parameters: the minimum frequency \( f_i \) from which the analysis starts and the number of channels to consider \( N_p \). Then, the QFT is computed as follows:

\[
X[k] = \sum_{n=0}^{N-1} w[k,n] \cdot x[n] \cdot e^{-j2\pi n/N} \quad \text{for } k=1\ldots N_p
\]

where \( w[k,n] \) indicates the Hamming windowing function: \( w[k,n] = \alpha + (1 - \alpha) \cos(2\pi n/N[k]) \) with \( \alpha = 25/46 \).

The number of temporal samples to consider for each channel is \( N[k] = \frac{SR \cdot Q}{f_k} \).

For example, assuming C4 (261.6 Hz) as the lowest note, the number of samples to consider is \( N[k]_{\text{max}} = N[1] = \frac{SR \cdot Q}{f_1} = \frac{44100 \cdot 34}{261.6} \approx 5732 \).

Once computing the QFT, the identification of sound fundamental frequency is made by a simple algorithm. Then, a FFT analysis starting from the calculated fundamental frequency allows to detect the amplitudes of the signal harmonics; their position may be found searching the local maximum around the frequencies that are integer multiples of the fundamental frequency.

Further parameters used for the classification are the absolute and the relative spectral centroids of the signal. The absolute spectral centroid is the first moment of the spectrum and can hence be calculated as

\[
C = \frac{\sum_k kE_k}{\sum_k E_k}
\]

where \( k \) is the generic frequency channel of the QFT and \( E_k \) the relative energy. The result is a frequency which should be characteristic for every instrument. The relative spectral centroid is the ratio between the absolute spectral centroid and the fundamental frequency \( f_i \) and is therefore defined as

\[
C_R = \frac{C}{f_i}
\]

The value of \( C_R \) is usually between 0.5 and 5. In Fig. 1, we show for example the spectra in the sustain phase of the same musical note (E4) produced by four traditional musical instruments: one stringed instrument (violin) and three wind instruments (flute, oboe, clarinet).

The spectra are obtained by using a windowing of 2048 time samples and a SR of 22050 Hz. The vertical lines appearing in the figure indicate the harmonics determined by the preprocessing procedure application.

The examination of the four spectra evidences that:
- for the flute, only the first harmonics are prominent;
- the spectrum of the oboe is different from the other three, since its initial harmonics are not the most important;
- in the case of clarinet, the second harmonic is nearly absent.

Fig. 1. FFT spectra of E4 note played, respectively, by violin, flute, oboe, and clarinet.

D. Luce and M. Clark show that these characteristics also hold in the case of other musical notes of the same instrument [9], contributing to identify it independently from the note played (and thus from the fundamental frequency). However, the spectrum is subject to large variations depending on the specific musician technique of playing.

### 3 Statistical Classifier

Commonly used classification systems can be placed into two principal categories: exclusive or nonexclusive. An exclusive classification system performs a partition of a feature space where each pattern (represented by a vector in this space) belongs to exactly one class. A nonexclusive...
classification system, instead, can assign a pattern to several classes. Such a method is hotly required in many real-world problems, where the pattern classes are frequently overlapping each other. From the nonexclusive point of view, overlapping regions contain a non-negligible information about the problem domain.

Statistical classifiers, based on Bayes’ theorem, are useful tools for obtaining the nonexclusive label of each pattern to be classified. Given a classification task of K classes, ω_1, ω_2, ..., ω_K, and an unknown pattern x, its label can be given by the K conditional probabilities P(ω_i | x), i = 1, 2, ..., K. Namely, each of them represents the probability of the class ω_i for generating the pattern x. The Bayes’ rule ensures the possibility for evaluating the probabilities P(ω_i | x), i = 1, 2, ..., K, by using both the class prior probabilities P(ω_i) and the class-conditional probability density functions p(x | ω_i); i.e.:

$$P(ω_i | x) = \frac{p(x | ω_i)P(ω_i)}{\sum_{j=1}^{K} p(x | ω_j)P(ω_j)}, \quad i = 1, 2, ..., K.$$  

In general, the classifier model is determined by means of a training set of examples. More precisely, both P(ω_i) and p(x | ω_i), i = 1, 2, ..., K, are estimated by considering each class ω_i independently (i.e., by considering in the training set only the patterns whose class is ω_i).

The class prior probability can be reasonably estimated by P(ω_i) ≡ N_i / N, where N is the overall number of patterns of the training set, and N_i is that of patterns belonging to the class ω_i. The class conditional density p(x | ω_i) can be estimated by using several techniques among those proposed in the technical literature.

The parameters of the clusters (mean, covariance and prior probability) can be estimated by applying the well-known expectation-maximization (EM) algorithm [6].

After the hierarchy is generated, the optimal clusters must be selected. According to how the brain perceives important structures of an image, only structures that survive over a large range of scale are perceived and therefore taken into consideration. When applied to clustering, the imitation of this mechanism requires the introduction of a scale parameter for simulating the scale of the training set, represented in the feature space, by an imaginary observer located at different distances [8].

Varying the scale parameter is similar to modify the distance from the data and, hence, the resolution level at which they are observed. In fact, in the generic step of the hierarchical algorithm, the cluster having the greatest scale (and thus the lower resolution level) is selected to be split. In conclusion, the most persistent clusters over the variation of the scale should be chosen. The best results are obtained by taking into account the intervals of existence of clusters present at different levels of the hierarchy (local stability) [8].

The splitting procedure of the cluster is based on a suitable choice among its principal directions (i.e., the eigenvectors of its covariance matrix). The patterns of the splitting cluster are divided in two groups (the offspring clusters) on the basis of their projections onto the chosen direction. Let h be the cluster to be split and ξ_h be the projection of its m-th pattern. The pattern is assigned to one group if ξ_h ≤ ξ_hu, where ξ_hu is the projection of the mean of cluster h; else it is assigned to the other group.

We define as scale of this cluster the Euclidean distance between the means of the resulting offspring clusters. Thus, when the scale is to be assigned to a cluster and the splitting direction is to be chosen, the splitting procedure is simulated for each principal direction. According to the scale-based criterion, the one scoring the greatest scale is chosen as the splitting direction.

4 Experiments and Conclusions

A recognition test was carried out for evaluating the performances of the classifier described in Sect. 3. Six instruments were considered: clarinet, flute, oboe, saxophone, piano and violin.

Several experiments were made, considering the 13 notes of the 4th octave from C4 (261.6 Hz) to C5 (522.2 Hz). In order to create the input patterns for the classifier, the following components are computed:

- the first 10 components correspond to the amplitude of the first harmonics, obtained with the procedure described in Sect. 2;
- the 11th component indicates the fundamental frequency of the note that was played;
- the 12th components contains the value of the absolute spectral centroid;
- the 13th component contains the value of the relative spectral centroid.

The considered spectra are determined by applying a FFT to a sequence of 4096 samples with a sampling frequency of 44100 Hz. The number of bits used in the A/D conversion is 16. For each instrument, 6 segments of each one of the 13 musical notes, included in the range from C4 to C5, are considered for the training set. Thus, the training set is constituted by 468 pattern, 78 for each instrument, concerning musical notes of the 4th octave randomly extracted from various musical executions.

The results reported in Tab. 1 show that the information added by the fundamental frequency and the spectral centroids improves the classification performances. The proposed classifier has been also tested by using another preprocessing method [12]. The corresponding classification error, obtained on the same training and test sets, is 18.79% with 38 clusters. That confirms the efficacy of our preprocessing method in this context.
<table>
<thead>
<tr>
<th>Input Features</th>
<th>Number of Features</th>
<th>Number of Clusters</th>
<th>Absolute Crisp Error (Test Set)</th>
<th>Crisp Error in % (Test Set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First 10 harmonics</td>
<td>10</td>
<td>28</td>
<td>19/390</td>
<td>4.87</td>
</tr>
<tr>
<td>First 10 harmonics + fundamental frequency</td>
<td>11</td>
<td>36</td>
<td>2/390</td>
<td>0.51</td>
</tr>
<tr>
<td>First 10 harmonics + fundamental frequency + spectral centroids</td>
<td>13</td>
<td>21</td>
<td>1/390</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Table 1. Results of the recognition experiments.

Finally, the proposed classification system favorably compares with the best performances obtained by other nonexclusive classifiers [4,5,7] based on neuro-fuzzy approaches as, for example, the well-known Simpson’s Min-Max model. In fact, they have been tested on the data sets as above: the resulting performances in terms of accuracy are worse, in the best case, of about 20%.

References: