Instrument Recognition Beyond Separate Notes -
Indexing Continuous Recordings

Arie A. Livshin, Xavier Rodet
livshin@ircam.fr, rod@ircam.fr
Ircam Centre Pompidou

Abstract
Some initial works have appeared that began to deal with
the complicated task of musical instrument recognition
in multi-instrumental music. Although quite a few papers have
already appeared on instrument recognition of single-
instrument musical phrases (“solos”), the work on solo
recognition is not yet exhausted. The knowledge of how to
deal well with solos can also help in recognition of multi-
instrumental music.

We present a process for recognition of a set of instruments
(bassoon, clarinet, flute, guitar, piano, cello and violin) in
solo recordings, which yields a high recognition rate.
Among the points that distinguish our work are a large and
very diverse solo database – 108 different solos, all by
different performers, which apparently supplies a good
generalization of the sound possibilities of each instrument,
and a large collection of features – 62 different feature
types. Using our GDE feature selection algorithm we
minimize the feature set and present the 20 features which
are most suitable for solo recognition in real-time, almost
without compromising the high recognition rate.
The paper ends by demonstrating that our real-time feature
set can also help performing instrument recognition in
“duet” music.

1 Introduction
The subject of musical instrument recognition in multi-
instrumental music (where several instruments play
concurrently) is difficult and is just beginning to get
explored (e.g. Eggink and Brown 2004). Quite a few papers
were written on instrument recognition in samples of
separate musical notes and some of them have also dealt
with recognition of single-instrument musical phrases, or
“solos” as we shall call them.

Dealing with solo performances is different and
apparently more complicated than dealing with separate
note databases, as the time evolution of each sound (attack,
sustain, release) is not well defined, the notes are not
separated, there are superpositions of concurrent sounds and
room echo, different combinations of playing techniques,
etc.

Marques and Moreno (1999) classified solos of 8 rather
diverse instruments, using one CD per instrument for
learning and one for classification. They compared 3 feature
types and 2 different classification algorithms, reaching a
70% recognition rate. Brown Houix and McAdams (2001)
classified 4 wind instruments using 4 different feature types
and achieved 82% recognition rate with the best parameters.
Martin (1999) classified sets of 6, 7 and 8 instruments,
reaching 82.3%, 77.9% and 73% recognition rates
respectively, using up to 3 different recordings of each
instrument. He used a relatively large feature set of 31 one-
dimensional features. For a comprehensive review of
instrument classification, see Herrera, Peeters and Dubnov
(2003).

The work on solo recognition is not yet exhausted.
Although the number of applications which in reality needs
solo recognition seems limited, yet the knowledge of how to
deal well with solos could also help in recognition of multi-
instrumental music, as will be demonstrated at the end of
this paper.

We begin by presenting a process for recognition of a set
of instruments (bassoon, clarinet, flute, guitar, piano, cello
and violin) which yields a high recognition rate. We are
using a large and very diverse solo database for training and
evaluation of the recognition process. This database holds
108 solo performances, all by different musicians, and
apparently supplies a good generalization of the different
sound possibilities of each instrument in various recording
conditions, playing techniques, etc., thus providing a good
generalization of the sounds each instrument is producing in
different recordings - what we call the “concept instrument”.
In order to evaluate the generalization ability of the
classifier, the same solos are never used both in the learning
and test sets; it was shown that classification evaluation
results where the training and learning sets both contain
samples recorded in very similar conditions can yield
misleading results (Livshin and Rodet 2003).

It is also one of the first times that such a large
collection of features is used for solo recognition – 62
different feature types (Peeters 2003) which were developed
and used in the Cuidado project. This diverse feature set
allows us, using our GDE feature selection algorithm, to
minimize the number of features and achieve a smaller
feature set best suited for solo recognition in real-time (of
our 7 instruments) with only a small compromise on the
recognition rate. We shall present these features, which
were actually implemented in a real-time solo recognition program.1

We end the paper by bringing results that show that the same descriptors and techniques we used for real-time solo recognition can also help in instrument recognition in duet performances.

2 Sound Database

In this paper we use 108 different real solo performances (by “solo” we mean that a single instrument is playing, in monophony or polyphony) of 7 instruments: bassoon, clarinet, flute, classical guitar, piano, cello and violin. These performances, which include classical and modern music, were taken from commercial CD’s (containing new or old recordings) and MP3 files played and recorded by professionals and amateurs.

Each solo was performed by a different musician and there are no solos which were taken from the same concert. In the evaluation process the same solo is never used both in the learning and in the test sets. The reason for these limitations is that we want the evaluation process to show the system’s ability to generalize - classify new musical phrases which were not learned and were recorded in different recording conditions, different instruments and played by different performers than the learning set.

It was shown that the evaluation results of a classification system which does learn and classify sounds performed on the same instrument and recorded in the same recording conditions, even if the actual notes are of a different pitch, are much higher than when classifying sounds recorded in different recording conditions. This is because such an evaluation process actually shows the system’s ability to learn and then recognize specific characteristics of specific recordings and not the ability of the system to generalize (Livshin and Rodet 2003).

2.1 Preprocessing

The solos were downsampled to 11Khz 16bit. If a solo recording was in stereo, arbitrarily, only the left channel was taken. A two minute piece was used from each solo recording and cut into 1 second windows with an overlap of 0.5 second – 240 cuts from every solo. The feature descriptors were computed on each 1 second solo-cut.

3 Feature Descriptors

The feature set used in this paper was written by Geoffroy Peeters as part of the Cuidado project. For a comprehensive explanation of all the features, see Peeters 2003.

Most of the features2 (except the ones computed on the whole signal) were computed using a sliding window of 60 ms with a step of 20 ms. For each solo-cut of 1 second, the average and standard deviation of these sliding windows over the 1 second range was used.

Initially, we have used a relatively large feature collection – 62 different features of the following types (Peeters and Rodet 2002):

Temporal Features. Features computed on the signal as a whole (without division into frames), e.g. log attack time, temporal decrease, effective duration.

Energy Features. Features referring to various energy content of the signal, e.g. total energy, harmonic energy, noise part energy.

Spectral Features. Features computed from the Short Time Fourier Transform (STFT) of the signal, e.g. spectral centroid, spectral spread, spectral skewness.

Harmonic Features. Features computed from the Sinusoidal Harmonic modeling of the signal, e.g. fundamental frequency, inharmonicity, odd to even ratio.

Perceptual Features. Features computed using a model of the human hearing process, e.g. mel frequency cepstral coefficients, loudness, sharpness.

Later in the paper we shall reduce the feature number by using the GDE feature selection algorithm.

4 “Minus-1 Solo” Evaluation Method

The feature data is normalized to the range of 0 - 1 (“min-max” normalization). Every solo in its turn is removed from the database and classified by the rest of the solos. This process is repeated for all solos, and the average grade for each instrument is reported along with the average grade among all instruments. This result is more informative than reporting the average grade per solo, as the number of solos for each instrument might be different.

The classification itself is done by first performing Linear Discriminant Analysis (LDA) (McLachlan 1992; Martin and Kim 1998) on the learning set, multiplying the resultant coefficient matrix with the test set and then classifying using the K Nearest Neighbors (KNN) algorithm. The “best” K is estimated from a range of 1-80 by using the leave-one-out (Livshin, Peeters and Rodet 2003) method on the learning set.3

1 Examples for possible applications are real-time indexing of recording studio tracks and real-time movie sound effects recognition for the hearing impaired.

2 Some features produce more than one value, e.g. the MFCC’s; We use the term “features” regardless of their number of values.

3 The “best K” (on average) for our database is estimated as 33 for the full feature set and 39 for the “real-time” set. Experiments with solo-cuts with an overlap of 0.75 second instead of 0.5 (resulting in 480 solo-cuts per solo instead of 240), reported a “best K” of 78 for the full feature set and 79 for the “real-time” set.
5 Feature Selection

In order to find the most important features we have used our Gradual Descriptor Elimination (GDE) feature selection method (Livshin, Peeters and Rodet 2003). GDE uses LDA repeatedly to find the descriptor which is the least significant and removes it. This process is repeated until no descriptors are left. At each stage the recognition rate of the system is estimated.

In this section we have set the goal to achieve a smaller feature set which will be quick to compute (allowing us to reach recognition in real-time) and will almost not compromise the recognition rate, compared to the complete feature set. By “real-time” we mean here that while the solo is recorded or played the features of each 1 second fraction of the music are computed and classified immediately after it was performed, before the following 1 second piece has finished playing/recording.

We removed the most time-consuming features and used GDE to reduce the feature-data until the number of features went down from 62 to 20. Using these features we have actually implemented a real-time solo phrase recognition program which works on a regular Intel Processor and is written in plain Matlab code (without compiling or integrating machine language boost routines). Naturally, this program computes the LDA matrix and finds the best K for integrating machine language boost routines. We can indeed see that the “real-time” average grade is rather close to the “complete set”. It is interesting to note that reducing the feature set has actually improved the recognition rate of the flute; LDA cannot always eliminate confusion caused by interfering features.

6 Results

Table 1 shows the Minus-1 Solo recognition results:

<table>
<thead>
<tr>
<th>Instrument</th>
<th>“Real-Time” 20 features</th>
<th>“Complete Set” 62 features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bassoon</td>
<td>86.25 %</td>
<td>90.24 %</td>
</tr>
<tr>
<td>Clarinet</td>
<td>79.29 %</td>
<td>86.93 %</td>
</tr>
<tr>
<td>Flute</td>
<td>83.33 %</td>
<td>80.87 %</td>
</tr>
<tr>
<td>Guitar</td>
<td>86.34 %</td>
<td>87.78 %</td>
</tr>
<tr>
<td>Piano</td>
<td>91.00 %</td>
<td>93.88 %</td>
</tr>
<tr>
<td>Cello</td>
<td>82.18 %</td>
<td>88.72 %</td>
</tr>
<tr>
<td>Violin</td>
<td>88.27 %</td>
<td>88.47 %</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>85.24 %</strong></td>
<td><strong>88.13 %</strong></td>
</tr>
</tbody>
</table>

Table 1. Average recognition rate of 1 second solo-cuts

We can indeed see that the “real-time” average grade is rather close to the “complete set”. It is interesting to note that reducing the feature set has actually improved the recognition rate of the flute; LDA cannot always eliminate confusion caused by interfering features.

6.1 Real-Time Feature Set

Table 2 shows the resulting 20 feature list for the real-time classification sorted by importance, from the most important feature to the least. We can see that the 10 most important features are the first 4 moments and the Spectral Slope, computed on both the perceptual and spectral models.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Perceptual Spectral Slope</td>
<td>83.8</td>
</tr>
<tr>
<td>2. Perceptual Spectral Centroid</td>
<td>50</td>
</tr>
<tr>
<td>3. Spectral Slope</td>
<td>45.2</td>
</tr>
<tr>
<td>4. Spectral Centroid</td>
<td>40</td>
</tr>
<tr>
<td>5. Spectral Kurtosis</td>
<td>37.5</td>
</tr>
<tr>
<td>6. Spectral Skewness</td>
<td>37.9</td>
</tr>
<tr>
<td>7. Perceptual Spectral Kurtosis</td>
<td>37.9</td>
</tr>
<tr>
<td>8. Spectral Kurtosis</td>
<td>25</td>
</tr>
<tr>
<td>9. Spectral Spread</td>
<td>25</td>
</tr>
<tr>
<td>10. Perceptual Spectral Skewness</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 2. A sorted list of the most important features for real-time solo classification (of our 7 musical instruments)

For a full explanation of every feature, see Peeters 2003.

7 Duet Examples

In Table 3 we give several selected examples for instrument recognition in real performance “duets” (where 2 instruments are playing concurrently) using our solo-recognition process with the real-time features. This section is not pretending to be an extensive research of duet classification, but comes to demonstrate that successful solo recognition might actually be useful for instrument recognition in multi-instrumental music.

<table>
<thead>
<tr>
<th>Musical Piece</th>
<th>Instrument #1</th>
<th>Instrument #2</th>
<th>Errors</th>
<th>Total correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castelnuovo: Sonatina</td>
<td>Bassoon 16.2</td>
<td>Piano 83.8</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Stockhausen: Tierkreis</td>
<td>Flute 50</td>
<td>Clarinet 50</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Seelis: Suite</td>
<td>Flute 71.1</td>
<td>Clarinet 28.9</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Carter: Esprit rude</td>
<td>Flute 45</td>
<td>Clarinet 52.5</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Kirchner: Triptych</td>
<td>Violin 37.5</td>
<td>Cello 60</td>
<td>97.5</td>
<td>97.5</td>
</tr>
<tr>
<td>Ravel: Sonata</td>
<td>Violin 38.5</td>
<td>Cello 59</td>
<td>97.5</td>
<td>97.5</td>
</tr>
<tr>
<td>Martinu: Duo</td>
<td>Violin 27</td>
<td>Cello 70.3</td>
<td>97.3</td>
<td>97.3</td>
</tr>
<tr>
<td>Pachelbel: Canon in D</td>
<td>Flute 17.6</td>
<td>Cello 77.5</td>
<td>95.1</td>
<td>95.1</td>
</tr>
<tr>
<td>Procaccini: Trois pieces</td>
<td>Bassoon 44.4</td>
<td>Piano 50</td>
<td>94.4</td>
<td>94.4</td>
</tr>
<tr>
<td>Bach: Cantata BWV</td>
<td>Flute 45.2</td>
<td>Cello 45.2</td>
<td>90.4</td>
<td>90.4</td>
</tr>
<tr>
<td>Sculptured: Fulfillment</td>
<td>Flute 25</td>
<td>Cello 63.9</td>
<td>88.9</td>
<td>88.9</td>
</tr>
<tr>
<td>Ohana: Flute duo</td>
<td>Flute 86.8</td>
<td>Cello 79.5</td>
<td>86.3</td>
<td>86.3</td>
</tr>
<tr>
<td>Bach: Cantata BWV</td>
<td>Violin 6.8</td>
<td>Cello 79.5</td>
<td>86.3</td>
<td>86.3</td>
</tr>
<tr>
<td>Pachelbel: Canon in D</td>
<td>Violin 0</td>
<td>Cello 84.6</td>
<td>84.6</td>
<td>84.6</td>
</tr>
<tr>
<td>Idrs: Aria</td>
<td>Bassoon 43.2</td>
<td>Flute 16.2</td>
<td>59.4</td>
<td>59.4</td>
</tr>
<tr>
<td>Feidman: Klezmer</td>
<td>Clarinet 40</td>
<td>Guitar 60</td>
<td>46.7</td>
<td>46.7</td>
</tr>
<tr>
<td>Copland: Sonata</td>
<td>Clarinet 45.5</td>
<td>Piano 25</td>
<td>45.5</td>
<td>45.5</td>
</tr>
<tr>
<td>Giuliani: Iglou</td>
<td>Violin 8.1</td>
<td>Cello 48.6</td>
<td>40.5</td>
<td>40.5</td>
</tr>
</tbody>
</table>

Table 3. Duet classification with our real-time solo recognition program
From each of the real performance duets, a 1 minute section was taken where both instruments are playing together and each second was classified by our real-time solo recognition program. The first column in Table 3 contains the partial name of the musical piece. The second and third columns show the percentage of solo-cuts which were actually classified as the two instruments that play the duet. The third column contains the misclassification percentage of solo-cuts and specifies which instruments were mistakenly “recognized”. The last column is the total percentage of solo-cuts that were correctly classified as one of the playing instruments.

We can see that there is a considerable number of examples where the classification was correct although, as we know, the classifier is very naïve and does not use f0 nor attempts any source separation. We shall study in future work why some instrument combinations produce specific recognition errors and how to prevent them, e.g. we can see that the guitar was a most common misclassification (in recognition errors and how to prevent them, e.g. we can see why some instrument co mbinations produce specific attempts any source separation. We shall study in future work why some instrument combinations produce specific recognition errors and how to prevent them, e.g. we can see that the guitar was a most common misclassification (in duets) and that we need features to discriminate it better.

8 Summary

In this paper we presented a process for continuous recognition of musical instruments in solo recordings which yields a high recognition rate. Our results are based on evaluation with a large and very diverse solo database which allowed us a wide generalization of the classification and evaluation processes, using the diverse sound possibilities of each instrument, recording conditions and playing techniques.

Using a big feature set and our GDE feature selection algorithm we considerably reduced the number of features, down to a feature set which allows us to perform real-time instrument recognition in solo performances. This small feature set delivers almost the same recognition rate as the complete set.

Lastly, we have shown that our recognition process and "real-time" feature set, without modifications, are also useful for instrument recognition in duet music. These results exemplify our initial claim that knowledge achieved in learning to deal well with solos could be also useful for instrument recognition in multi-instrumental performances.

9 Future work

We will continue working on instrument recognition in multi-instrumental music. We intend to study the reasons for correct recognition in some duets and incorrect recognition in others by our solo classifier. We have started working on a multi-instrument recognition process where each solo-cut could be classified into more than one instrument. This process also provides a confidence level for every classification.

We will work on partial “source separation”, where we shall not attempt to actually separate the instruments but rather to weaken the volume of some of the tones and then use a modified “solo” classifier.

We shall perform feature selection among more features, some of which will be especially designed with multi-instrumental recognition in mind.

Acknowledgments

Many thanks to Geoffroy Peeters for contributing his feature computation routines and sharing his extensive knowledge and experience.

Thanks to Emmanuel Vincent for sharing his solo database.

References


