Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

The importance of the non-harmonic residual for automatic musical instrument recognition of pitched instruments

Arie Livshin[1] and Xavier Rodet[2]

[1] IRCAM Centre Pompidou, 1 place Stravinsky, Paris 75004, France
arie.livshin@ircam.fr

[2] IRCAM Centre Pompidou, 1 place Stravinsky, Paris 75004, France
xavier.rodet@ircam.fr

ABSTRACT

In different papers dealing with automatic musical instrument recognition of pitched instruments, the features used for classification are based solely on the fundamental frequencies and the harmonic series, ignoring the non-harmonic residual. In this paper we explore whether instrument recognition rate of pitched instruments is decreased by removing the non-harmonic information present in the sound signal.

1. INTRODUCTION

Musical instruments with definite pitch are usually based on a periodic oscillator such as a string or a column of air with non-linear excitation. In consequence, the sound of pitched instruments is mostly composed of a harmonic series of sinusoidal partials, i.e. frequencies which are integer multiples of the fundamental frequency *(f0)*.

Various articles dealing with instrument recognition of pitched instruments (as opposed to drums, for example) use solely sound feature descriptors computed on the harmonic series of the signal for classification of the sounds (see [1] for example). However, if we subtract the harmonic series from the original sound there is a non-harmonic residual element left. This residual is far from being 'white noise'; it is heavily filtered by the nature of the instrument itself as well as the playing technique, e.g. scraping noises (guitar), breathing (flute), etc., and may contain inharmonic sinusoidal partials as well as non-sinusoidal 'noise'. This residual, although much less prominent than the harmonic series, might still be useful for instrument recognition.

In this paper we explore how using feature descriptors based only on the harmonic series compares with using the whole signal for musical instrument recognition. In order to perform this comparison as 'fairly' as possible,

we extract from pitched musical instrument samples the harmonic series information, i.e. the *f0*s, harmonic partials and corresponding energy levels, and using additive synthesis reproduce the signals directly from the harmonic series, thus creating synthesized 'images' of the original signals which lack any non-harmonic information.

Next, we compute the same set of feature descriptors on both the original sound samples and the synthesized ones. The original and synthesized sound groups are divided separately into training and test sets and instrument classification is performed on both groups independently. Finally, the instrument recognition results of the original and synthesized sound groups are presented and compared.

There is a practical motivation for this experiment for the field of instrument recognition; when performing instrument recognition in multi-instrumental, polyphonic music, it is quite difficult as well as computationally expensive [2] to perform full source-separation and restore the original signals from the polyphonic mix in order to recognize them separately, while estimating the harmonics of the separate notes is a relatively easier task. See [3] for example of instrument recognition in polyphonic recordings, where the estimated harmonic series is used for performing semi-source-separation technique called 'Source Reduction'.

2. DATABASE

Our sound database consists of 4823 monophonic samples of single notes of 10 'musical instruments': Bassoon, Clarinet, Flute, Trombone, Trumpet, Bass, Bass pizzicato, Violin, Violin pizzicato and Piano. The pizzicatos are considered separate instruments as they are very different from the sounds of the bowed strings.

The sound samples were collected from 11 different sound sample databases, providing sounds from different recording conditions and instruments. The set of all the samples of a specific instrument taken from a single database (e.g. all the violin samples from collection 'X'), will be referred to as an 'instrument instance'. The total number of instrument instances is 71. All the sounds are sampled in 44Khz, 16bit, mono.

3. RESYNTHESIS

In order to remove the residual signal, all the samples were analyzed and resynthesized using high precision parameters with the additive analysis/synthesis program 'Additive' [4]. As the note name and octave of each sample are provided by the sound databases, very precise additive analysis/synthesis parameters were tailored depending on the given *f0*s. For example, the analysis/synthesis window size was set to 4*(*f0* - (one tone)) allowing vibrato, FFT size to (4 * next power of 2 above the window size), and many others.
[Figure 1] exemplifies a resynthesized trumpet note. We can see that the residual remaining after subtracting the resynthesized signal from the original, is quite low.

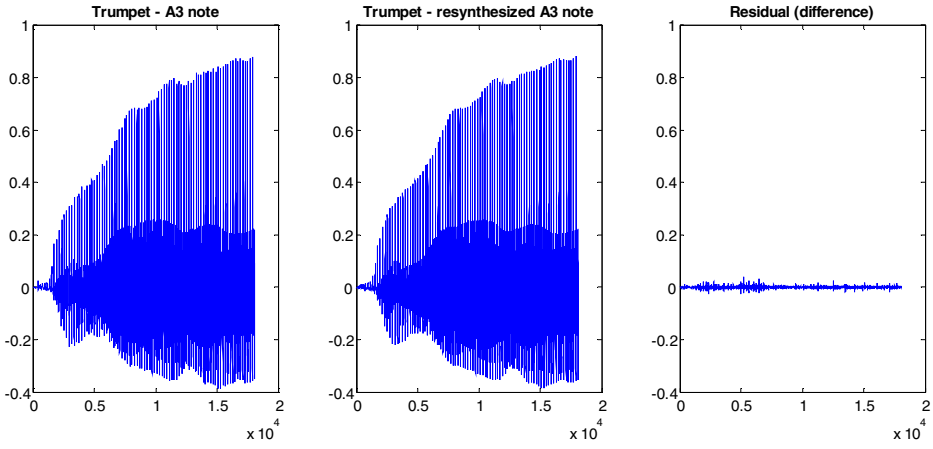

Figure 1. Left to right: original trumpet sample (A3 note), resynthesized sample, the residual

4. FEATURE DESCRIPTORS

The same feature set[1] is computed on both the original sample set and the resynthesized one. Except several features which were computed using the whole signal, most of the features were computed using a sliding frame of 60 ms. with a 66% overlap. For each sample, the average and standard deviation of these frames were used by the classifier.

In order to encapsulate the various characteristics of the signals, the feature set we use is quite large and includes 62 different feature types:

4.1. Temporal Features

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

4.2. Energy Features

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

4.3. Spectral Features

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

4.4. Harmonic Features

Features computed from the Sinusoidal Harmonic modeling of the signal, e.g. *f0*, inharmonicity, odd to even ratio.

4.5. Perceptual Features

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

After computation, the feature descriptors are normalized to the range of [0..1] using Min-Max Normalization.

5. CLASSIFICATION

Instrument recognition is performed on the original and resynthesized sets of samples separately. In order to get meaningful instrument recognition results it is necessary to use independent data in the learning and test sets [6]. For this purpose, we introduce the 'Minus-1 Instance' evaluation method: each instrument instance is removed in its turn from our joined database[2] and its samples are classified by all the remaining ones, thus no sound samples from the same individual instrument and recording conditions are present in both the learning and test sets. The average recognition rate per instrument, over all the samples of each instance, is reported.

Each classification phase begins by computing Linear Discriminant Analysis (LDA) transformation matrix using the learning set. LDA reduces the dimensionality of data with c classes down to c-1 dimensions, while maximizing between-class scatter (i.e. distance between the means of different classes) and minimizing within-class scatter (i.e. variance inside each class) by maximizing the Fisher criterion [7].

After dimension reduction, the test set is classified by the learning set using the K-Nearest-Neighbors (KNN) classifier. Different K values in the range of [1..80] are tested at each classification phase and after the whole Minus-1 Instance classification process completes, the best K for the whole classification process is reported.

6. RESULTS

The average Minus-1 Instance recognition rate per instrument for the original samples is 93.94% (using K=14 for KNN). This recognition rate is quite high compared to other papers dealing with instrument recognition of separate notes using independent learning and test sets (see [6] and [8] for example), and as our database is relatively large and diverse, emphasizes the intuitive claim that enriching the learning database with

[1] The feature computation routines were written by Geoffroy Peeters as part of the Cuidado project. Full details on these features can be found in [5].

[2] Reminder: our database is joined from samples originating from 11 different sound databases. At each classification step, the samples of one instrument from one of these databases are removed.

samples recorded in different recording conditions improves its generalization power [6].

Looking at the "original samples" confusion matrix [Table 1], we see that the highest misclassification rates are: 10.48% of the samples of the clarinet instances are misclassified as flute; the flute in its turn has 'lost' 4.85% to the clarinet; the trombone 'lost' 6.35% to the bassoon and 5.4% to the trumpet.

The average Minus-1 Instance recognition rate per instrument for the resynthesized samples is 89.69% (using K=6 for KNN). This recognition rate is only 4.25% lower than the average recognition rate using the original samples, and is still relatively high.

Comparing the confusion matrices of the resynthesized samples [Table 2] to the original ones [Table 1], we can see that the recognition rate of almost all the instruments worsened somewhat. The most noticeable declines: the violin pizzicato went down by extra 11.75% compared with the original samples, losing a total of 8.86% to the piano and 5.61% to the bass pizzicato; the trombone went down by 10.57%, losing a total of 12.79% to the bassoon and 9.53% to the trumpet. The clarinet went down additional 6.31%, losing a total of 10.99% to the flute and 5.27% to the trumpet; the trumpet went down another 5.61% and the bassoon another 3.84%.

	Bassoon	**Clarinet**	**Flute**	**Trombone**	**Trumpet**	**Bass**	**Bass pizz.**	**Violin**	**Violin pizz.**	**Piano**
bassoon	**98.40**	0.64	0.64	0.00	0.32	0.00	0.00	0.00	0.00	0.00
Clarinet	0.00	**85.85**	10.48	0.00	1.39	1.44	0.00	0.84	0.00	0.00
Flute	0.00	4.85	**92.86**	0.00	0.52	0.00	0.00	1.77	0.00	0.00
Trombone	6.35	0.00	0.00	**86.25**	5.40	0.00	0.00	2.00	0.00	0.00
Trumpet	0.69	1.25	1.84	0.41	**95.50**	0.23	0.00	0.00	0.07	0.00
Bass	0.40	0.05	0.00	0.00	0.00	**97.84**	0.00	1.71	0.00	0.00
Bass p.	0.00	0.00	0.00	0.00	0.00	0.25	**98.06**	0.00	0.00	1.69
Violin	0.14	0.14	2.34	0.00	0.00	0.86	0.00	**92.68**	2.88	0.96
Violin p.	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.11	**97.12**	1.86
Piano	1.63	0.00	0.00	0.00	0.00	2.02	0.28	0.00	1.26	**94.81**

Table 1. Confusion matrix (rows classified as columns) of 'Minus-1 instance' using original samples

	Bassoon	**Clarinet**	**Flute**	**Trombone**	**Trumpet**	**Bass**	**Bass pizz.**	**Violin**	**Violin pizz.**	**Piano**
bassoon	**94.56**	0.27	1.08	1.67	0.27	2.14	0.00	0.00	0.00	0.00
Clarinet	0.99	**79.54**	10.99	0.00	5.27	1.18	0.00	2.02	0.00	0.00
Flute	0.00	5.28	**90.89**	0.00	0.17	0.58	0.00	3.03	0.00	0.06
Trombone	12.79	0.00	0.00	**75.68**	9.53	2.00	0.00	0.00	0.00	0.00
Trumpet	0.69	1.61	3.83	1.98	**89.89**	0.64	0.00	1.30	0.07	0.00
Bass	0.54	0.40	0.29	0.00	0.00	**96.04**	0.00	2.73	0.00	0.00
Bass p.	0.00	0.00	0.00	0.00	0.00	0.20	**97.84**	0.00	1.41	0.55
Violin	0.04	0.69	3.04	0.00	0.62	2.15	0.00	**93.37**	0.00	0.09
Violin p.	0.05	0.11	0.00	0.00	0.00	0.00	5.61	0.00	**85.37**	8.86
Piano	0.90	0.19	0.11	0.06	0.11	0.39	1.28	0.00	3.25	**93.71**

Table 2. Confusion matrix (rows classified as columns) of 'Minus-1 instance' using resynthesized samples

7. CONCLUSIONS

We can see that removing the residual has somewhat blurred the differences between 'similar' instruments, for example, removing the hammer strike sound from the piano and the plucking sound from the pizzicatos has increased the classification confusion between these groups of samples, which have similar temporal envelope; removing the transient has also 'mellowed'

the bassoon sound, increasing the confusion with the trombone, etc.

To summarize: although the recognition rate has decreased noticeably for some instruments, reaching a maximum of 12.79% loss for the trombone, it is important to notice that the average grade of the resynthesized sounds was reduced by only 4.25% and is still quite high – 89.7%. These results show that using only the harmonic information may indeed be enough for achieving rather good instrument recognition rates, although it does cause some distinguishing information loss.

8. FUTURE WORK

The recognition rate could be improved further by specifically addressing the weakest instruments in this paper, namely the Trombone and Clarinet, which already had a lower recognition rate than the other instruments when the original samples were classified. Adding specifically tailored descriptors to deal with these instruments and enriching the database even more, can help classify them better.

We have seen that using only the harmonic series does not considerably lower the average instrument recognition rate although some instruments suffer more than others. This means that instrument recognition in polyphonic, multi-instrumental music could indeed be performed with rather high results without performing full source-separation[3]; using multiple *f0* estimation algorithms (like the one in [10]), estimated harmonic partials could be classified with an instrument classifier without losing too much distinguishing information.

9. ACKNOWLEDGEMENTS

This work is partly supported by the "ACI Masse de données" project "Music Discover".

Thanks to Geoffroy Peeters for using his feature computation routines.

[3] There are also papers that research performing instrument recognition in multi-instrumental, polyphonic music, without using source-separation. These compute the feature descriptors directly on the mixed signal; see [10] for example.

10. REFERENCES

[1] J. Eggink, G. J. Brown, "Instrument recognition in accompanied sonatas and concertos," Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2004, pp. 217-220

[2] E. Vincent, X. Rodet, "Instrument identification in solo and ensemble music using Independent Subspace Analysis," Proc. International Conference on Music Information Retrieval (ISMIR), 2004

[3] A. Livshin, X. Rodet, "Indexing Continuous Recordings," Proc. of 7'th international conference on Digital Audio Effects (DAFx), Naples, Italy, 2004, pp. 222-227

[4] "IRCAM Real time applications - Additive and HMM".URL:http://www.ircam.fr/58.html?L=1&tx_ircam_pi2%5BshowUid%5D=35&ext=2

[5] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," 2003. URL:http://www.ircam.fr/anasyn/peeters/ARTICLES/ Peeters_2003_cuidadoaudiofeatures.pdf

[6] A. Livshin, X. Rodet, "The Importance of Cross Database Evaluation in Musical Instrument Sound Classification," Proc. International Symposium on Music Information Retrieval (ISMIR), 2003

[7] G. J. McLachlan, "Discriminant Analysis and Statistical Pattern Recognition". New York, NY: Wiley Interscience, 1992.

[8] Marques, J. "An Automatic Annotation System for Audio Data Containing Music". Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999

[9] S. Essid, G. Richard, B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," IEEE Transactions on Speech and Audio Processing, 2006, to be published.

[10] C. Yeh, A. Röbel, X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, 2005.