INSTRUMENT IDENTIFICATION IN POLYPHONIC MUSIC SIGNALS BASED ON INDIVIDUAL PARTIALS

*Jayme Garcia Arnal Barbedo**

State University of Campinas
DECOM/FEEC
Campinas, SP, Brazil

George Tzanetakis

University of Victoria
Department of Computer Science
Victoria, BC, Canada

ABSTRACT

A new approach to instrument identification based on individual partials is presented. It makes identification possible even when the concurrently played instrument sounds have a high degree of spectral overlapping. A pairwise comparison scheme which emphasizes the specific differences between each pair of instruments is used for classification. Finally, the proposed method only requires a single note from each instrument to perform the classification. If more than one partial is available the resulting multiple classification decisions can be summarized to further improve instrument identification for the whole signal. Encouraging classification results have been obtained in the identification of four instruments (saxophone, piano, violin and guitar).

Index Terms— instrument identification, polyphonic musical signals, pairwise comparison, partial-based classification

1. INTRODUCTION

The information of which instruments are present in a given musical signal can benefit a number of applications in the audio signal processing area. For example, the accuracy of methods trying to classify musical signals into genres can be greatly improved by such information. Also, sound source separation techniques can be better tuned to deal with the characteristics of particular instruments.

A number of techniques to identify musical instruments have been proposed in the literature. Many proposals can only deal with monophonic signals (e.g. [1]). The monophonic case is in general less challenging than the polyphonic one, as there is no interference among instruments. Some proposals can deal with solo phrases with accompaniment (e.g. [2]). In those cases, normally the solo instrument has to be strongly dominant, so the signals characteristics are quasi-monophonic. Also, methods based on solo phrases analysis normally provide a classification for the whole signal, instead of individual notes. There are also some proposals that were designed specifically to deal with instrument duets (e.g. [3]). Methods capable of dealing with polyphonies containing up to four instruments were also proposed [4, 5, 6, 7], but all of them have some important limitations: the possible instrument combinations are set a priori, limiting the generality [4]; only notes with duration above 300 ms are processed [5]; the accuracy for higher polyphonies (four or more instruments) is below 50% [6, 7].

This paper presents a simple and reliable strategy to identify instruments in polyphonic musical signals that overcomes many of the main limitations faced by its predecessors. The identification is based on a pairwise comparison between instruments. A related but more complex approach was used in [4]. This pairwise comparison approach has some particular characteristics:

- Whenever possible, the features used to represent each pair of instruments were designed to characterize individual partials instead of individual notes. Using individual partials is desirable because, in western music, most content of each instrument overlaps with the contents of other sources both in time and frequency, making the identification a challenging task. However, one can expect to find at least some unaffected partials throughout the signal, which can be explored to provide cues about the respective instrument. As a consequence, the proposed algorithm can identify the instruments even when all but one partial collide with partials belonging to other instruments and 2) if more than one isolated partial is available, the results can be combined to provide a more accurate identification. Among the pairs of instruments considered in this paper, only one (piano-guitar) could not be represented in a partial basis. However, the feature adopted in this particular case allows correct identification even if all partials of the instrument suffer interference from other instruments, as will be seen in the following sections.

- Only one feature is used to linearly separate two instruments. A comparison of this approach with traditional machine learning methods is presented in Section 3.

As a result of the characteristics mentioned above, the proposed strategy is able to deal with polyphonies of any order, provided that at least one partial of each instrument does not suffer interference from other instruments.

Summarizing, this paper aims to show that: accurate estimates are possible using only a small number of partials, without requiring the entire note; integrating over partials can potentially improve the results; using only one carefully designed feature to separate each pair of instruments is more effective than traditional machine learning methods.

To evaluate the effectiveness of this approach we consider four instruments - alto saxophone, violin, piano and acoustic guitar - that were carefully selected to represent the more general situations expected when more instruments are considered. More information about this choice can be found in Section 2.

2. THE METHOD

2.1. Experimental setup

The 1,000 mixtures used in the training stage were artificially generated by summing individual notes from four instruments taken from the RWC database [8]. Half the mixtures contain two instruments,

*Thanks to Foreign Affairs and International Trade Canada for funding. The author performed the work while at the Department of Computer Science, University of Victoria, Canada.

Table 1. Separation accuracy for each pair of instruments.

	SV	SP	SG	PV	VG	PG1	PG2
Acc.	90%	93%	94%	90%	93%	95%	83%

and the other half contain three instruments. The instruments and respective fundamental frequencies that compose each signal were taken randomly, provided that each instrument has at least one isolated partial. The entire note range of each instrument is considered, and the signals are sampled at 44.1 kHz. The 1,000 mixtures used in the tests were generated in the same way but, in order to provide some cross-database validation, half of them contain at least one instrument from the University of Iowa musical instrument samples database [9]. Also, none of the instrument samples used in the trainig set was used in the test set.

Some actions must be performed in order to make the instrument identification possible. First, the signal must be segmented according to the onsets of the notes, which can be performed automatically using tools like those presented in [10]. Segments that are smaller than 100 ms - which occur in less than 5% of the cases - are not considered. After that, the number of instruments present in each segment must be determined, using, for example, the technique described in [11]. Finally, the fundamental frequency (F0) of each instrument must be estimated (e.g. [12]). In this work, all the information regarding segmentation, number of instruments and fundamental frequency in each segment is assumed to be known, in order to avoid cascading errors.

2.2. Feature selection and extraction

As commented before, each specific pair of instruments has a feature associated with it. Therefore, the features to be extracted depend directly on which instruments are being considered.

Four instruments - alto saxophone (S), violin (V), piano (P) and acoustic guitar (G) - were chosen to validate the strategy. Hence, there are six possible pairs of instruments. The instruments in the pairs SP, SG, VP and VG have considerably different characteristics (temporal waveform, spectral content, etc), while the instruments in the pair SV have some similar characteristics and the instruments in the pair PG are closely related, as discussed in [13]. In this way, the technique can be tested under different levels of difficulty.

In total, 54 features were considered. Most of the features were implemented based on [14] and [15], slightly modified to be extracted on individual partials - a partial, in the context of this work, refers to a narrow spectral band in which a harmonic of a given fundamental frequency lies. The feature selection for each pair of instruments aimed for the best linear separation between the instruments present in the training dataset – such a separation is given by a single boundary value that separates the value ranges of each instrument. . All selected features are calculated individually for each partial, except that adopted for the pair PG, as will be seen in the following. Table 1 shows the best separation accuracy achieved for each pair. PG1 and PG2 refer to the main and alternative features of pair PG.

Table 1 was generated using the following features:

SV: bandwidth containing 90% of the energy of the partial. This feature is calculated according to the following steps:

- This feature must be calculated in the frequency domain, thus a discrete Fourier transform is calculated for the segments, and then the magnitude spectrum is extracted.

- The value of this feature for partial n is given by the narrowest band that concentrates 90% of the energy of the frequency interval $[(n-0.4)\cdot F0, (n+0.4)\cdot F0]$.

- If there is a partial from another instrument in that interval, the band $[(q-0.1\cdot f, q+0.1\cdot f)]$ is removed from the calculation, where q is the center frequency of the interfering partial, and f is the fundamental frequency associated to q.

SP: center of gravity of the temporal envelope. This feature is calculated according to the following steps:

- Each partial is isolated by a third-order Butterworth filter, according to the interval $[(n-0.4)\cdot F0, (n+0.4)\cdot F0]$.

- The absolute value of the Hilbert transform, followed by a Butterworth filter, is used to estimate the time envelope of the partial.

- The envelope center of gravity is calculated according to

$$c_n = \frac{1}{T}\cdot\sum_{t=1}^{T} t\cdot x_n(t), \tag{1}$$

where x is the temporal envelope, t is the time index, T is the number of samples, and n is the partial index.

SG, PV, VG: note skewness. It measures the asymmetry of the envelope around the c_n. It is calculated according to

$$sk_n = \sum_{t=1}^{T} x_n(t)\cdot\left(\frac{t}{T}-c_n\right)^3 \Big/ sp_n^{1.5}, \tag{2}$$

where sp_n is the note spread, which is given by

$$sp_n = \sum_{t=1}^{T} x_n(t)\cdot\left(\frac{t}{T}-c_n\right)^2. \tag{3}$$

PG: dominance of the 100-120 Hz band. No feature calculated for individual partials was able to reliably separate this pair. However, it was observed that all guitars in the database generate a peak somewhere in the band 100-120 Hz due to body resonances, which in general does not happen for piano. This has motivated the creation of a new feature which is calculated according to

$$d = M/L, \tag{4}$$

where M is the mean of the magnitude spectrum in the 10-100 Hz band, and L is the magnitude of the strongest peak in the 100-120 Hz band. The smaller is the resulting value, the more dominant is the peak. If the next value larger than the peak is closer than 30 Hz, d is multiplied by two. This aims to compensate for the cases in which the peak in the 100-120 Hz is mainly due to a neighbor partial.

A partial may be located in the 100-120 Hz band, in which case this feature is ineffective. If this happens, an alternative feature, related to the main one, is calculated. It was observed that both the piano and the acoustic guitar generate a small peak at 49 Hz, which tends to be slightly more prominent for the guitar. It is calculated in the same way as Eq. 4, but here M is the mean magnitude spectrum in the bands 44-48 Hz and 50-54 Hz, and L is the magnitude of the strongest peak in the 47-51 Hz band.

The peak located at 49 Hz is present in all tested piano and guitar samples. Since the samples come from different instruments and databases, it seems that this phenomenon is not due to specificities of the instruments bodies or the acoustics of the room. However, a definitive answer will only be possible after further investigation with additional samples from other databases.

2.3. Instrument identification procedure

As commented before, the algorithm assumes that the number of instruments and respective fundamental frequencies are known. Therefore, it is possible to identify the isolated partials that do not

suffer interference from other instruments. Only isolated partials are submitted to the next steps of the algorithm.

Each isolated partial is submitted to the pairwise comparison, in which an instrument is chosen as the winner for each pair, and the partial is labeled according to the instrument with most wins. If a two-way equality occurs, the winner is the instrument that won when both instruments were considered in a pair. If a three-way equality occurs, the winner is the instrument whose related features have the greatest overall identification accuracy.

The same procedure is repeated for all partials related to that fundamental frequency. Then, the instrument with more wins throughout the isolated partials is taken as the correct one for that note. There are two criteria to break possible equalities. The first one just takes into account the total number of wins that each instrument got when each partial was considered separately. If the equality remains, the instrument assigned to the strongest partial is taken as final winner. The same procedure is repeated for all fundamental frequencies, until all instruments have been identified.

3. EXPERIMENTAL RESULTS

Fig. 1 shows six confusion matrices, each one considering a different number of isolated partials (number in the top-left of the matrices). An average of 750 instances were used for each instrument.

Tests were also performed with the alternative feature described in Section 2.2. The results were very close to those shown in Fig. 1, except in the case of the acoustic guitar, for which the accuracy dropped to 82.1%, no matter the number of partials. However, the alternative feature will be used only in the few cases with very low fundamental frequencies, thus having little impact in the overall performance of the approach.

Some conclusions can be drawn from Fig. 1. First, it can be observed that the results are good even when only one partial is available, with an overall accuracy close to 91%. Also, as expected, the results improve as more partials are available, until 96% accuracy is achieved when six or more isolated partials are available. Such results indicate that the pairwise comparison approach is indeed effective, because the overall accuracy is greater than the accuracies obtained for each pair of instruments using individual features.

In a first analysis, the results may seem unrealistically good. Three factors explain those results. First, one must take into account that only four instruments are considered. Also, all instruments are taken from the same database, which can cause slightly biased results [16]. Future tests are expected to include a cross-validation among databases. Finally, a very effective feature was found to the difficult pair PG, which significantly improved the overall results.

As commented in Section 2.1, part of the samples used in the tests come from the University of Iowa [9] database. The results considering only such samples are very close to those shown in Fig. 1 – the accuracy dropped, in average, 1.8%.

It is also instructive to know how the partial-wise approach compares to cases in which the whole notes are considered instead. Table 2 shows the comparison. The results for the whole note approach were obtained by filtering out partials known to be colliding, and then averaging the individual features over the entire note. The P and N in the first line refer to partial-wise and note approaches, and the numbers indicate the number of available partials. The results are in percentage of correct identifications.

As can be seen, for violin the performance is clearly poorer when the entire note is used. This is because the significant differences among the partials make the partial-wise approach more suitable, as those differences are individually taken into account, and not averaged over the entire note.

Table 2. Comparison between partial and note approaches.

	P1	N1	P3	N3	P6	N6
Sax	89.2	88.8	92.3	92.0	95.1	94.4
Piano	91.1	91.0	94.9	94.0	96.7	96.3
Violin	90.0	87.3	93.8	90.1	95.1	91.4
Guitar	93.3	93.3	95.2	95.0	97.7	97.1

Table 3. Effect of backward onset misplacement.

Common Partials	20%	40%	60%	80%	100%
1/3	0.92	0.88	0.85	0.83	0.82
1/2	0.86	0.82	0.80	0.75	0.74
1	0.53	0.47	0.39	0.32	0.28

Tests were also performed to determine if using individual features is advantageous when compared to classical machine learning methods, which in general combine several features to perform the classification. The overall accuracies using support vector machines (SVM) and k-nearest neighbors (KNN) were, respectively, about 5% and 7% worse than the results shown in Fig. 1. Both the SVM and KNN were trained with the same features presented in Section 2.2. Although the results may vary significantly, it was observed that, in general, if it is possible to find features capable of separate each pair of instruments with an accuracy of at least 90%, the use of individual partials is advantageous. Deeper and more detailed tests to clarify this matter are planned to be carried out in the near future.

Because the complete system would have to rely on the information provided by an onset detector and a fundamental frequency estimator, it is useful to analyse how the inclusion of such tools would influence the results. Regarding the onset detector, three kinds of errors may occur: a) errors smaller than 10% of the frame length, which have little impact in the accuracy of the instrument identification, because the characteristics of the partials are only slightly altered; b) large errors, with the estimated onset placed after the actual position, which have little effect over sustained instruments, but may cause problems for instruments whose notes decay over time because the main content of the note may be lost – in those cases, the instrument identification accuracy drops almost linearly with the onset error – for example, a 30% forward error in the onset position will result in 30% less accurate estimates for instruments with decaying notes; c) large errors, with estimated onset placed before the actual position, in which case a part of the signal that does not contain the new note is considered, and the severeness of the effects are linked to the number of common partials between the spurious and actual notes, and to the length of the onset displacement. Table 3 shows the effects of the backward onset misplacements (given in percentage of the actual frame) when the interfering note has 1/3, 1/2 and all partials in common with the note to be classified. The values given in Table 3 represent the relative accuracy given by A_e/A_i, where A_e is the accuracy of the method when there is onset misplacement, and A_i is the accuracy of the method when the onset is in the correct position. As can be seen, when there are some partials that are not affected by the interfering note, the method is able to compensate in part the problems caused by the onset misplacement.

As stated before, the results showed here refer to the first phase of development of the algorithm. The next steps of research will deal with a much greater number of instruments (at least 25), which will bring some new challenges. The procedure described in this paper is able to compensate, to a certain degree, the flaws that may occur when using single features to separate pairs of instruments. As

1	sax	piano	violin	guitar
sax	89.2	3.3	6.3	1.2
piano	1.9	91.1	6.1	0.9
violin	6.4	3.2	90.0	0.4
guitar	2.0	1.8	2.9	93.3

2	sax	piano	violin	guitar
sax	91.0	2.8	5.1	1.1
piano	1.7	93.2	4.2	0.9
violin	5.0	2.6	92.1	0.3
guitar	1.6	1.5	2.4	94.5

3	sax	piano	violin	guitar
sax	92.3	2.7	4.3	0.7
piano	1.6	94.9	2.6	0.9
violin	3.9	2.3	93.8	0.0
guitar	1.5	1.2	2.1	95.2

4	sax	piano	violin	guitar
sax	93.4	2.4	3.7	0.5
piano	1.4	96.0	2.3	0.3
violin	3.2	2.1	94.7	0.0
guitar	1.1	0.9	1.5	96.5

5	sax	piano	violin	guitar
sax	94.3	1.9	3.4	0.4
piano	1.1	96.4	2.3	0.2
violin	3.1	2.0	94.9	0.0
guitar	0.8	0.5	1.3	97.4

6	sax	piano	violin	guitar
sax	95.1	1.6	3.3	0.0
piano	1.0	96.7	2.1	0.2
violin	2.9	2.0	95.1	0.0
guitar	0.8	0.3	1.2	97.7

Fig. 1. Confusion matrices using the features described in Section 2.2. The numbers in the top-left of the matrices indicate the number of isolated partials available. The instruments in the first column represent the targets, and those in the first row represent the actual classification.

new instruments are introduced, it is expected that some pairs will be similar to the point it may not be possible to find a feature able to perform the separation accurately. In such cases, the ability of the pairwise comparison to compensate isolated inaccuracies will be tested to the limit. On the other hand, a greater number of instruments will result in a much greater number of pairs. This means that the weight of each pair in the final classification will be greatly reduced, and so will be the influence of problematic pairs. Although it is not possible to predict how the algorithm will perform under such conditions, good results using the pairwise comparison approach with a large number of classes have been achieved before in the context of music genre classification [17]. Although the problems of music classification and instrument identification are somewhat different, this previous success provides some evidence that the pairwise approach may also be successfully extended to the identification of a large number of instruments.

4. CONCLUSIONS

This paper presented the first results obtained in the development of a new approach to identify musical instruments in polyphonic musical signals. It employs a pairwise comparison approach, and the identification, in its most basic level, is mostly performed in an individual partial basis. The winner instrument for a given note is finally identified after the results are summarized from the partial level to the note level. Results show that the strategy is robust and accurate.

The next steps of development will include a greater number of instruments, which will bring new challenges that will test the algorithm to the limit. Future tests will also include signals from other databases, as a cross-validation between databases may provide further significance to the results.

5. REFERENCES

[1] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," *EURASIP J. Applied Signal Process.*, vol. 2003, pp. 5–14, 2003.

[2] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 174–186, 2009.

[3] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proc. of the IEEE*, vol. 92, pp. 712–729, 2004.

[4] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, pp. 68–80, 2006.

[5] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," *EURASIP J. Applied Signal Process.*, vol. 2007, 2007.

[6] P. Leveau, D. Sodoyer, and L. Daudet, "Automatic instrument recognition in a polyphonic mixture using sparse representations," in *Proc. Int. Conf. on Music Inf. Retrieval*, 2007.

[7] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering," in *Proc. Int. Conf. on Music Inf. Retrieval*, 2007.

[8] M. Goto, "Development of the RWC music database," in *Proc. Int. Cong. Acoustics*, 2004, pp. 553–556.

[9] "University of Iowa musical instrument samples database," http://theremin.music.uiowa.edu/.

[10] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 1035–1047, 2005.

[11] J. G. A. Barbedo, A. Lopes, and P. J. Wolfe, "Empirical methods to determine the number of sources in single-channel musical signals," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 1435–1444, 2009.

[12] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 804–816, 2003.

[13] D. Fragoulis, C. Papaodysseus, M. Exarhos, G. Roussopoulos, T. Panagopoulos, and D. Kamarotos, "Automated classification of piano-guitar notes," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, pp. 1040–1050, 2006.

[14] A. Eronen, "Automatic musical instrument recognition," M.S. thesis, Tampere University of Technology, Finland, 2001.

[15] M. R. Every, "Discriminating between pitched sources in music audio," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, pp. 267–277, 2008.

[16] A. Livshin and X. Rodet, "The importance of cross database evaluation in sound classification," in *Proc. Int. Conf. on Music Inf. Retrieval*, 2003.

[17] J. G. A. Barbedo and A. Lopes, "Automatic musical genre classification using a flexible approach," *J. Audio. Eng. Soc.*, vol. 56, pp. 560–568, 2008.