Citation:

Somerville, P and Uitdenbogerd, A 2007, 'Note-based segmentation and hierarchy in the classification of digital musical instruments', in Proceedings of the 2007 International Computer Music Conference, Copenhagen, Denmark, 27-31 August 2007, pp. 240-247.

NOTE-BASED SEGMENTATION AND HIERARCHY IN THE CLASSIFICATION OF DIGITAL MUSICAL INSTRUMENTS

Peter Somerville and Alexandra L. Uitdenbogerd
School of Computer Science and Information Technology
RMIT University
Melbourne, Vic., Australia, 3000

ABSTRACT

The ability to automatically identify the musical instruments occurring in a recorded piece of music has important uses for various music-related applications. This paper examines the case of instrument classification where the raw data consists of musical phrases performed on digital instruments from eight instrument families. We compare the use of extracted features from a continuous sample of approximately one second, to the use of a systematic segmentation of the audio on note boundaries and using multiple, aligned note samples as input to classifiers. The accuracy of the segmented approach was greater than the one of the unsegmented approach. The best method was using a two-tiered hierarchical method which performed slightly better than the single-tiered flat approach. The best performing instrument category was woodwind, with an accuracy of 94% for the segmented approach, using the Bayesian network classifier. Distinguishing different types of pianos was difficult for all classifiers, with the segmented approach yielding an accuracy of 56%. For humans, broadly similar results were found, in that pianos were difficult to distinguish, along with woodwind and solo string instruments. However there was no symmetry between human comparisons of identical instruments and different instruments, with half of the broad instrument categories having widely different accuracies for the two cases.

1. INTRODUCTION

There are many ways in which users engage with a digital music collection. As these collections continue to grow in size and popularity, an increased range of methods for finding music automatically is also likely to be required. Such a method is the location of music based on the types of musical instrument found in the recording. The instruments used are also likely to be useful predictors of whether someone is likely to prefer a given piece of music. For example, consider a person's aversion to piano accordion music. Other related uses of instrument identification include the management of digital sounds used by a musician, and the automatic labelling of segments of a long recording for studio processing.

A major objective of this research is to determine the best techniques for an application that allows for songs to be retrieved based on an instrument's timbre. In our previous work [6] we addressed the classification of musical instruments, where only a single note sample was provided as input. In this paper we consider the case where there are multiple notes within each sample, varying in speed, volume and melodic shape across the collection of data. This paper provides evidence for whether the automated classification of digital musical instruments is more successful using an unsegmented or a segmented approach, where segmentation is on note boundaries. *Monophonic* music (one note occurring at a time) was generally used in the experiments.

We also compare human perception of musical instrument timbre with automated classification techniques. Unlike nearly all published work on instrument classification, we focus on virtual, software and synthesized instrument types for our data collection.

Feature extractors used in the experiments were: Spectral Centroid, Spectral Rolloff, Spectral Flux, Zerocrossings, RMS (Root Mean Square - amplitude envelope) and Mel-Frequency Cepstral Coefficients (MFCC). The classifiers applied to the instrument samples were decision trees (J48), K-nearest neighbor (KNN), Naive Bayes and Bayesian Networks (BayesNet).

The best performing classifier was BayesNet where the segmented approach returned an overall classification accuracy average of 77%. This percentage referred to the fine-grained in isolation instrument classification experiments. When broad instrument classification was undertaken, the result was 68%. Fine grained instrument classificiation in isolation refers to comparing instruments within the same instrument family. For example, classifying all pianos within the piano category. Broad instrument classification refers to pianos being compared to organs and woodwind and to all the other categories

When comparing a two-tiered approach to one where classification occurred into 52 instrument categories, the former worked best, but it only attained an accuracy of 54%.

All three approaches, segmented, non-segmented and human-based classification had difficulty in distinguishing pianos. The woodwind instruments returned the best results for the segmented approach, but was more difficult to distinguish by human subjects, whilst the lead synthesizers category was best for unsegmented classification.

The paper is broken into the following sections. Related work is followed by the Approach used, Data Sources, Experiments and Results finishing with the Discussion, Conclusion and suggestion of work to follow.

2. RELATED WORK

While we were unable to find work dealing with the automatic classification of synthetic instruments, there are some papers that address instrument classification using multi-note or polyphonic recordings. We discuss some of these here. Table 1 summarizes authors, feature selectors used and classification techniques applied for all the cited references.

Essid et. al. [2] addressed the issue of instrument recognition in polyphonic music (multiple notes occurring at the same time), by representing combinations of instruments that are likely to be played together with respect to a certain musical genre. The jazz genre was used and sound excerpts from commercial recordings were used. Ensembles using a combination of the following ten instruments were used in the experiments: double bass, drums, piano, percussion, trumpet, tenor sax, electro-acoustic guitar, Spanish guitar and male and female singing voices.

The results showed that by using a hierarchical classification algorithm, the recognition of classes consisting of combinations of instruments was possible. The scheme produced a hierarchy of nested clusterings. The approach started with the same number of clusters as classes and then measured the distances between pairs of clusters. The closest pairs were then grouped into new clusters. This process was continued until all classes lay in a single cluster. The work showed an average accuracy of 53% for segmented music with respect to the instruments played [2].

Sandvold et. al. [4] used feature-based modelling for classifying percussive sounds mixed in polyphonic music. Localised sound models were built for each recording using features and combined with prior knowledge (general models) to improve percussion classification. Categories were kick, snare, cymbal, kick+cymbal, snare+cymbal and not-percussion. The results returned an accuracy of values 20% higher than that of general models.

Simmermacher et. al. [5] presented a study on classifying musical instruments occurring in solo passages of classical music recordings. Segments from concertos and sonata pieces were collected in order to distinguish trumpet, flute, violin and piano in solo passages. The training and tests set included different recordings of the four instruments, with all except the piano samples having background accompaniment. The researchers achieved an accuracy of about 94% for their best classifier.

Agostini et. al. [1] looked at the problem of the recognition of musical instruments from monophonic musical signals. The research focused on the extraction of score-like attributes from an audio signal, which included the notes and their durations and sound-source recognition. A dataset of over 1007 tones from 27 musical instruments was used. Grouping were made based on instrument family or pizzicato/sustained nature of the sounds. The pizzicato category contained piano and related instruments, rock strings and pizz. strings and the sustained category contained strings, woodwinds and brass. Some of the features which require further explanation include inharmonicity which is related to the difference between the frequencies of the overtones of a fundamental sound and the whole number multiples of the fundamental's frequency, and, harmonic energy skewness which combines inharmonicity with the energy confined in each partial. Inharmonicity, spectral centroid and the energy contained in the first partial were the most relevant features. One of the classifiers used was discriminant analysis which endeavors to look for combinations of variables which best explain the data. It also attempts to model the difference between the classes of data. Support vector machines and quadratic discriminant analysis provided a classification success rate of close to 70%. The string instrument family was difficult to classify whilst satisfactory results were possible with the brass and woodwind families.

3. APPROACH USED

We discuss below the fact that all the experiments involving automatic instrument classification used the same basic apparatus. We also describe the feature extraction for our two main techniques.

The experiments conducted in this research involved extracting features from digital instrument samples and then classifying them into instrument categories. We used two main approaches that we call *segmented* and *unsegmented*. Figure 1 explains the unsegmented and segmented approaches and the steps needed from the feature selection to the classification stage. In the unsegmented approach, approximately one second of the sound file was used, leaving the file intact and treating the file as it is.

3.1. Segmentation

The segmented approach examined the same audio file as the unsegmented approach, but sliced it into smaller audio files based on the detection of the beginning of notes in the sample. The tool used was Beatcreator, which can apply segmentation to audio files based on note onset detection. The result was numerous chunks of smaller audio files that contain at most, a couple of consecutive notes. In most cases the smaller files contained just one note. Once the audio files were sliced, the first five files greater than 18 kB in size (and 200 milliseconds(ms) in duration) were used for each instrument sample. The data for the five audio files were placed alongside one another resulting in an equivalent amount of audio data to that of the unsegmented approach. Both approaches examined a total audio length of approximately 1 second duration.

In Beatcreator, a sensitivity of 7 and a threshold of -41dB was used for all slicing. Sensitivity is measured on a scale of 0 to 8, with high values leading to more events or notes being detected. The threshold setting allowed filter-

Authors	Features	Classifiers	Training/Test Set Size	Classes
Essid et. al. [2]	**Temporal** - Autocorrelation coefficients, Zero crossing rates, Local temporal waveform moments, Amplitude modulation features Cepstral - Mel-Frequency Cepstral Coefficients (MFCC) **Spectral** - first two coefficients of an Auto-regressive analysis, Spectral centroid, Spectral width, Spectral asymmetry, Spectral kurtosis(peakedness/flatness), MPEG-7 which provides spectrum flatness, Spectral slope, Frequency derivative of the Constant-Q coefficients **Perceptual** - covers sharpness and spread of the sound	- SVM(Support Vector Machine) - GMM(Gaussian Mixture Models)	1000/500	13
Sandvold et. al. [4]	Correlation-based feature selection algorithm	- K-Nearest Neighbors with k=1	1136/1419	6
Simmermacher et. al. [5]	**Temporal - perception based** - Zero Crossing Rate, Root Mean Square, Spectral centroid and Flux **Spectral - MPEG-7 based** - included 7 of a possible 18 features, Harmonic centroid, Harmonic deviation, Harmonic spread,Harmonic variation, Log-attack-time, Temporal centroid, and Spectral centroid Mel-Frequency Cepstral Coefficients (MFCC)	- k-NN - multilayer perceptron (MLP)(feedforward based Neural network), - Support Vector Machine	1160/800	4
Agostini et. al. [1]	Spectral centroid, Spectral bandwidth, Inharmonicity, Harmonic energy skewness, Zero crossing rate	- Discriminant analysis, Quadratic discriminant analysis - Canonical discriminant analysis, K-Nearest Neighbours, Support Vector Machines	1007	27

Table 1. The Feature Selectors and Classifiers used in Cited Work on Instrument Classification.

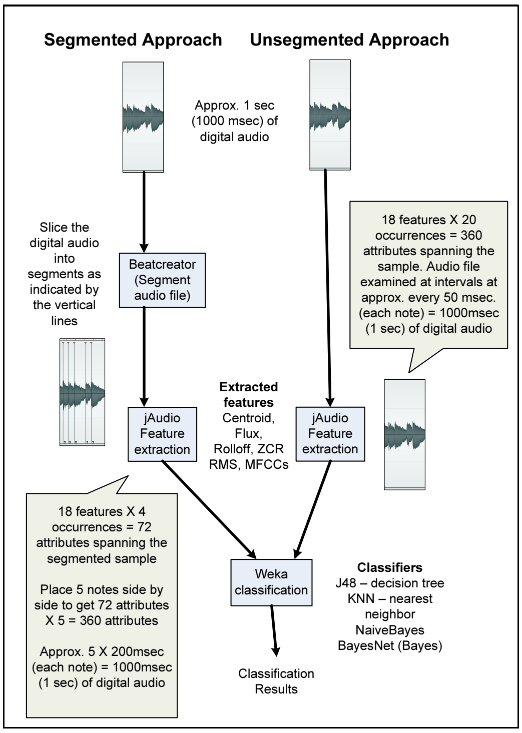
Segmented Approach
Unsegmented Approach
Approx. 1 sec (1000 msec) of digital audio
Slice the digital audio into segments as indicated by the vertical lines
Beatcreator (Segment audio file)
18 features X 20 occurrences = 360 attributes spanning the sample. Audio file examined at intervals at approx. every 50 msec. (each note) = 1000msec (1 sec) of digital audio
Extracted features Centroid, Flux, Rolloff, ZCR RMS, MFCCs
jAudio Feature extraction
jAudio Feature extraction
18 features X 4 occurrences = 72 attributes spanning the segmented sample
Place 5 notes side by side to get 72 attributes X 5 = 360 attributes
Approx. 5 X 200msec (each note) = 1000msec (1 sec) of digital audio
Weka classification
Classifiers
J48 – decision tree
KNN – nearest neighbor
NaiveBayes
BayesNet (Bayes)
Classification Results

Figure 1. Audio feature extraction and classification method for segmented and unsegmented approaches. The number of features and attribute counts are provided in calculations.

ing out of unwanted events below a given level. Threshold ranged from -60dB to 0dB where a setting of -60dB will find many events and 0dB will find very few. Unfortunately, further details about the these parameters are unavailable.

3.2. Feature Extraction

The feature extractors used for both segmented and unsegmented approaches were Spectral Centroid, Spectral Flux, Spectral Rolloff, Zero Crossing, RMS and 13 MFCCs. ACE's jAudio, an open source package was used for the feature extraction tasks [3].

Using the unsegmented audio file, each instrument sample examined contained 20 occurrences of the 18 features, giving 360 total attributes spanning approximately 1000 ms of digital audio.

The segmented approach used the same feature extractors as the unsegmented approach. For each smaller segmented audio file, there were 4 occurrences of the 18 features spanning the entire sample. This gave 72 attributes for each segmented sample. Placing 5 notes or 5 segmented samples side by side gave 360 attributes in total, being the same as the unsegmented approach.

For each audio file, both approaches had features extracted every 50 ms. Sound files were stored as 44.1 kHz, 16 bits, mono digital audio files. Within jAudio, a window size of 2048 samples was used.

3.3. Classification

Our experiments used Decision Trees (J48), KNN, BayesNet and NaiveBayes as implemented by the data mining software Weka, using default values. Evaluation was based on ten-fold stratified cross-validation. The stratified approach is where Weka attempts to properly represent each instrument class in both training and test sets.

4. DATA SOURCES

The data sources for our experiments came from software and hardware based instruments rather than sampling from real acoustic instruments. These included hardware synthesizers (Korg Trinity and Yamaha CS2X), digital piano (Roland EP85) and software based VST instruments. Software based instruments also include soundfonts which are sample banks originally designed for Sound Blaster soundcards but now can be used in many virtual sampling software packages. The eight categories of instruments we used were brass, plucked, organs, pianos, woodwind, solo string, synthesizer and pad (noted background synth&pad in the following), and lead synthesizers (noted leadsynth in the following). Every family contained six different instruments except for background synth&pad which contained ten. Background synth&pad used ten instruments as it is a broad category containing a wide range of possible instruments. The plucked category contained instruments such as harp, harpsichord and nylon string guitar. Solostring differed to background synth&pad in that the former included instruments such as a solo violin and solo cello whereas the latter contained big pad sounds and string ensembles.

Drums, along with similar percussive instruments and human singing voices were not used in these experiments. All instruments and groupings can be seen in Figure 2.

The experiments were based on midi files that were converted into digital audio files based on the different instruments chosen. For our experiments we used three sets of music with different characteristics. The first set consisted of the following instrument categories: brass, plucked, organs, pianos, woodwind and solostring and all used digital audio derived from the same set of midi files. They came from the following six pieces of music: Tchaikovsky - Swanlake - prelude, BirdLand full band, Handel's Water Music, a Reggae piece, a Hardrock piece and a Latin piece. The second set, the background synth&pad music category, used digital samples generated from midi files that comprised many notes played simultaneously. The third set, the leadsynth category, used midi files that comprised short segments of quickly played lead lines. The final two sets used midi files created by the first author.

The following experiments cover broad and fine grained instrument classification and human timbre sensitivity aspects. Broad instrument classification experiments used 292 instances whereas the fine grained experiments used 36 for all instrument categories except for background synth&pad. The amount of 292 comprised 7*36 + 40 where the first calculation refers to all instruments except for the background synth&pad category, and the last number refers to the background synth&pad number of instances. The value of 40 is calculated from 10 instruments each having 4 different sound files originally generated from the midi files.

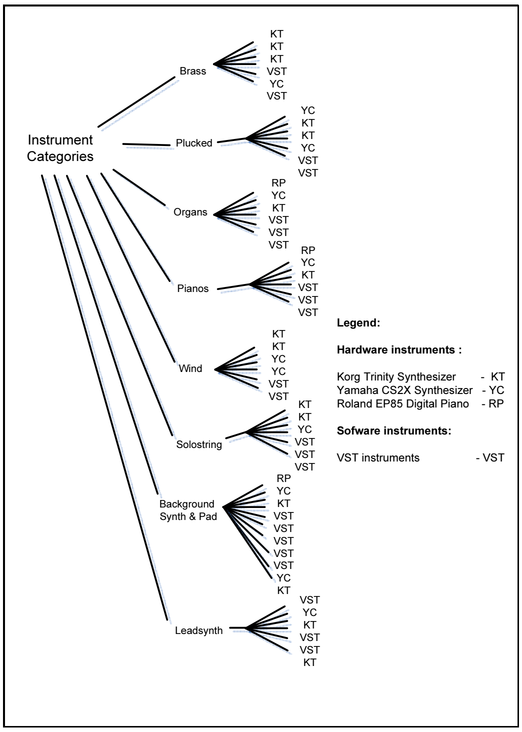

Figure 2. Instrument tree showing instrument categories and types. Hardware and software based instruments are included in the eight different instrument categories.

5. EXPERIMENTS

With our experiments, we aimed to learn more about the distinguishability of different electronic musical instrument timbres, both by humans and machine. Unlike our earlier work, which was based on single note samples [6], here we considered the case of monotimbral musical excerpts.

The first two experiments explore several variables for classification of musical instruments by timbre. In particular we compare the use of segmentation based on note boundaries with a simple unsegmented audio sample approach. We also compare the effects of using a two-level hierarchy with a flat classification structure. The final experiment examines human timbre perception for our instrument timbre data set.

5.1. Broad Instrument Classification

This experiment helps to answer the question of whether the automated classification of digital musical instruments was more successful using an unsegmented or segmented approach, where the classifier categories were pianos, organs, solostring, brass, woodwind, leadsynth, plucked and background synth&pads.

When segmentation is undertaken, the BayesNet classifier performed the best with a 68% average value for broad instrument classification. The BayesNet classifier also performed the best for the unsegmented approach with an average of 58%. The features that were most effective in the experiments were MFCCs. The instrument category with the highest average classification was background synth&pads with 84% whilst the poorest performing category was woodwind with 49%.

As can be seen in the confusion matrices in Figures 3 and 4, the brass and woodwind instrument categories were hard to distinguish for both the unsegmented and segmented methods. Values in the confusion matrices are shown as percentages.

```
BayesNet

   a     b     c     d     e     f     g     h   <-- classified as
 83.3  11.1   2.8   0.0   0.0   0.0   0.0   0.0  | a
  8.3  66.7   5.6   5.6   5.6   5.6   2.8   0.0  | b
 11.1  13.9  22.2  33.3   8.3   0.0   0.0  11.1  | c = brass
 13.9   8.3  27.8  30.6  16.7   2.8   0.0   0.0  | d = woodwind
 13.9   2.8   2.8  13.9  55.6   5.6   0.0   5.6  | e
 27.8  11.1   2.8   5.6   2.8  36.1   5.6   8.3  | f
  5.0  10.0   2.5   5.0   5.0   7.5  52.5  12.5  | g
  0.0   2.8   2.8   0.0   5.6   5.6   5.6  77.8  | h

  Legend:
         a = piano        e = solostring
         b = organ        f = plucked
         c = brass        g = bgsynandpad
         d = (wood)wind   h = leadsynth
```

Figure 3. Confusion matrix for the best performing broad instrument classifier using unsegmented audio data. The brass and woodwind instrument categories are highlighted as these instruments indicate the most confusion.

```
BayesNet

  a      b     c     d     e     f      g     h   <-- classified as
 83.3   8.3   0.0   0.0   8.3   0.0   0.0   0.0  | a
  8.3  66.7   0.0   8.3   5.6   2.8   0.0   8.3  | b
  0.0   5.6  61.1  16.7   8.3   2.8   0.0   5.6  | c = brass
  2.8   5.6  25.0  58.3   0.0   2.8   0.0   5.6  | d = woodwind
  5.6   0.0   5.6   2.8  75.0   5.6   0.0   5.6  | e
 22.2  13.9   2.8   5.6   0.0  47.2   2.8   5.6  | f
  0.0   2.5   0.0   0.0   0.0   7.5  82.5   7.5  | g
  0.0   0.0  11.1   2.8   2.8   5.6   8.3  69.4  | h


  Legend:
         a = piano        e = solostring
         b = organ        f = plucked
         c = brass        g = bgsynandpad
         d = (wood)wind   h = leadsynth
```

Figure 4. Confusion matrix for the best performing broad instrument classifer using segmentation.

5.2. Fine-Grained Instrument Classification

The following testing involves fine-grained instrument classes where the instruments are compared in isolation within instrument groups. The fine grained tests then extended to include all 52 instruments as separate classes. This is compared to a two-tiered classification approach.

5.2.1. *Testing within Instrument Families*

In this experiment we tested each broad instrument category separately, with classification being into specific instrument sounds within the category. For example, the woodwind instruments were classified into one of six specific classes in isolation from other instruments such as pianos.

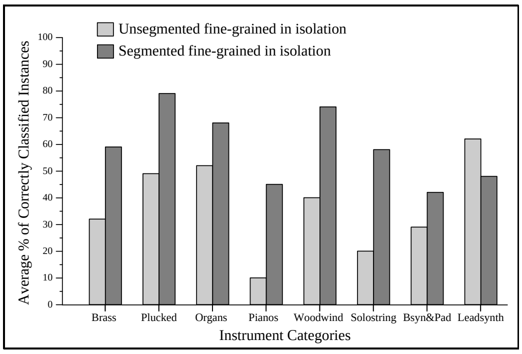

Figure 5. Classification averages taken across the four classifiers for the fine-grained instrument groups. Data for the unsegmented and segmented approaches are given.

```
J48

   a      b      c      d      e      f   <-- classified as
 33.3   16.7    0.0   16.7    0.0   33.3  | a
 16.7    0.0   66.7    0.0    0.0   16.7  | b = ep85piano2
 16.7   33.3   33.3    0.0   16.7    0.0  | c = kontaktscc1
 16.7    0.0    0.0   83.3    0.0    0.0  | d
  0.0    0.0    0.0    0.0   83.3   16.7  | e
 33.3   33.3    0.0    0.0   33.3    0.0  | f



BayesNet

  a       b      c       d      e      f   <-- classified as
 33.3   16.7    0.0     0.0    0.0   50.0  | a
  0.0   66.7   16.7     0.0   16.7    0.0  | b = ep85piano2
  0.0   50.0   16.7     0.0   33.3    0.0  | c = kontaktscc1
  0.0    0.0    0.0   100.0    0.0    0.0  | d
  0.0   16.7   16.7     0.0   66.7    0.0  | e
 33.3    0.0    0.0     0.0   16.7   50.0  | f


  Legend:
          a = cs2xwired           d = korgpa127isntitgrand
          b = ep85piano2          e = STAGrand2
          c = kontaktscc1         f = vstthegrand
```

Figure 6. Confusion matrices for two segmentation-based piano classifiers.

Averages of correctly classified instances for the different instrument categories for both the unsegmented and segmented approaches can be seen in Figure 5. The results were calculated by averaging across the individual results gained from the four classifiers. Classification for the unsegmented approach was generally quite poor. For the segmented approach, the plucked, woodwind and organ instrument categories returned the highest classification results. They were 79%, 74% and 68% respectively with pianos and background synth&pad being the lowest returning 45% and 42%. Pianos were also difficult to classify in the experiments undertaken in [6] where fewer instrument categories were used. Examining the piano results more closely indicated that there were generally two pianos that were confused with one another (see Figure 6). The first was a digital piano sound, and the second, a very clear and crisp sounding soundfont created in software.

When segmentation was conducted, the BayesNet classifier performed the best with a 77% average classification across all instrument categories. BayesNet was also the best performing classifier for the unsegmented approach but it only returned an average of 51%.

Using the unsegmented approach, the Spectral Centroid and Spectral Rolloff did not stand out as significant

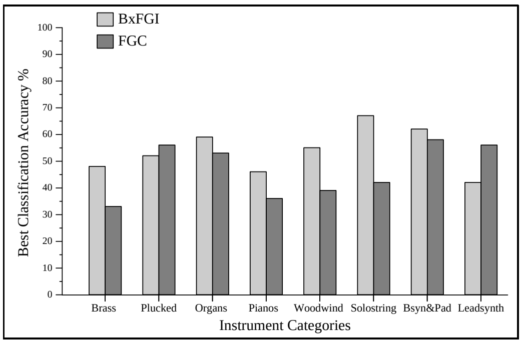

Figure 7. Accuracy of the two-tiered classifier (BxFGI) and flat classifier (FGC) for each instrument category.

features, whereas using the segmented approach, no feature performed poorly. Feature selectors which were significant for both approaches were RMS, Zero Crossings and MFCCs.

5.2.2. *Flat Versus Two-Tiered Instrument Classification*

When attempting to classify instances into a large set of categories, a hierarchical approach is often used. For example Essid and al. [2] used a hierarchical clustering algorithm to recognize classes consisting of combinations of instruments played simultaneously. We compare a two-tiered classification consisting of eight broad classes in the first tier with direct classification into 52 classes. The instruments and corresponding groupings can be seen in Figure 2. The best performing classifier was identified and used for each calculation.

We evaluated the two-tiered classification approach by multiplying the broad classifier success rate (B) by the best fine grained instrument in isolation of each type (FGI). These classification percentages were then compared with the success rate of the best fine-grained classifier (FGC). To generate the FGC percentage for individual fine-grained instrument classes, the 52 classes are grouped into 8 instrument categories where averages are calculated for each group. These percentages (FGC) are then compared to the ones resulting from the BxFGI calculations.

Figure 7 compares the success of classification within each class for the two-tiered and flat classifiers.

In most cases, the results in Figure 7 for the two-tiered classifier were better than for the flat classifier. The only exceptions were the plucked and leadsynth categories. It seems that improvements can occur using a hierarchical approach, even if the broad classifier is not perfect. Averaging across the eight instrument categories, the rates BxFGI and FGC were 54% and 47% respectively.

5.3. Human versus Machine

To gain an understanding of how well humans perceive differences in the timbre of digital musical instruments we gave web-based instrument classification surveys to participants with sufficient musical background and experience. Eight categories out of nine given to the participants involved selecting similar instruments within the same instrument category. The other category labelled ‘combined’ included a mixture of all instruments from each category.

Participants were asked some general questions about their musical skill level, instruments played and whether they had sound engineering and/or sound designing skills, or, were computer music composers. The number of years of experience was also recorded for the appropriate questions. In each instrument category, 20 pairs of instrument samples were presented. The duration of each sample was 3.5 seconds which could be played any number of times. The instrument pairs were randomly chosen from the relevant instrument categories and the ordering of the pairs was altered for half of the participants completing the survey.

Table 2 shows the number of choices that can be answered correctly as ’Yes’ or ’No’ in each of the instrument categories. The survey entry for woodwind, for instance, has nine instrument pairs whose sounds come from the same instrument and eleven instrument pairs whose sounds are from different instruments. The variation was given across instrument categories so that participants did not assume and expect the same number of ‘Yes’ and ‘No’ pairs. Participants were asked to decide whether the samples from the pair were from the same instrument. They selected ‘yes’ for pairs that sounded like they came from the same instrument and ‘no’, if they did not. Participants were asked to ignore in their judgments the volume, pitch (which included instruments played at different octaves), tempo and the tune played. Sixteen participants completed the survey, most of whom had extensive musical backgrounds, especially in the area of performance. The remaining participants were recruited from specialist mailing lists such as ACMA (Australasian Computer Music Association) and MUSIC-IR (Musical Information Retrieval).

5.3.1. *Survey Results*

To enable a comparison of results between the automated segmented approach and the human timbre perception approach, it was also necessary to determine the percentage of correctly classifed instances for the survey based approach. The average duration of completing the survey was one hour. The results were derived for each instrument category by calculating the average of the number of responses correctly answered as ‘yes’ and number of responses correctly answered as ‘no’.

Ignoring the combined category, which performed very well, the leadsynth and organ categories performed the best with average correct classification percentages of 79% and 77% respectively. Pianos (59%) and solostring (61%) had the lowest respective correct classification averages. For all categories except pianos, it was easier to correctly identify that two samples were from different instruments than from the same.

When two sounds came from the same instrument, participants had more difficulty with the brass, solostring

Instrument Category	*'Yes' answer tally*	*'No' answer tally*
Brass	10	10
Plucked	9	11
Organs	11	9
Pianos	9	11
Woodwind	9	11
Solostring	11	9
Bgd synth&pad	11	9
Leadsynth	10	10

Table 2. Indicates the weighting of 'Yes' and 'No' answers provided for participants on the survey for each instrument category.

Rating	*Ranking*	*Automated (segmented) approach*	*Human timbre perception*
Best	1	Woodwind	Leadsynth
	2	Plucked	Organs
	3	Solostring	Plucked
	4	Organs	Brass
	5	Brass	Bgd synth&pad
	6	Bgd synth&pad	Woodwind
	7	Leadsynth	Solostring
Poorest	8	Pianos	Pianos

Table 3. Comparison and overall ranking of instrument groups based on automated (segmented) classification and human timbre perception. Fine-Grained Instrument Classification results have been used for the automated method.

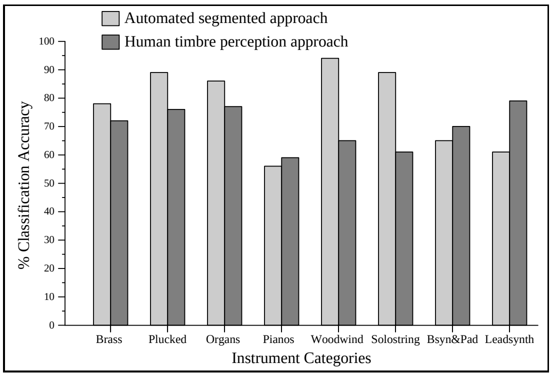

Figure 8. Automated segmented results versus Human timbre sensitivity results for all instrument categories. Each instrument category shows the percentage of correctly classified instances.

and woodwind than other categories. Two of these categories represent instruments into which the player has to blow. When two different instruments were presented, pianos and solostrings were the two classes that were the most difficult, having average classification percentages of 57% and 72% respectively. So for pianos, in approximately half the cases, participants thought two piano samples came from the same instrument when in fact, they did not.

We were not able to control the audio quality with our experiment, which could have affected the ability of participants to distinguish very similar timbres.

However, one of the participants completed the survey using average standard laptop speakers, and then later she used much better desktop speakers. In the second case, the results were generally better, especially when two samples came from the same instrument. However, the participant more frequently decided that samples from different instruments were from the same instrument.

One particular comment made by the participants during the surveys was: "It was very difficult to compare the likeness of two instruments when they were played at different ends of the keyboard spectrum". This is to be expected as the timbre of an instrument can change considerably across different octaves.

To compare the automated segmented method and the human timbre perception approach, Table 3 and Figure 8 have been provided. Table 3 shows a ranking from best to worst based on the instrument classes used in our experiments. The ranking for the automated results was derived from the fine-grained instrument classification segmented results. This ranking was based on the best percentage chosen from the results of the four classifiers. For the human-based experiments, the ranking positions of each instrument were derived from averaging the percentages of the selections correctly answered as ‘Yes’ or ‘No’.

6. DISCUSSION

Instruments within the piano category were difficult to classify no matter whether software or humans were used to test the experiments. One reason for this could be that the instruments within the piano family are more similar to one another than the instruments within any other family. For example, the woodwind instrument family includes flutes, oboes, clarinets, recorders and bassoons. The brass family has trumpets, french horns, and trombones whereas the piano category has different types of pianos. Other keyboard related instruments such as electric pianos and Rhodes style pianos were not used in the experiments. The classifiers, however, had more difficulty in classifying the leadsynth category. The reverse is true for woodwind. The plucked and organ instrument categories were easier to classify than other categories, no matter which approach was used. Confusion was evident between the woodwind and brass categories when using automated classification approaches. This was also difficult for human subjects, when deciding whether two samples from the same instrument were indeed from the same instrument.

As it was much easier for humans than machines to distinguish sounds from the leadsynth category, there is scope for more work to improve classification for such sounds. There has been little research other than ours on classification of sounds like these, which bear little resemblance to traditional acoustic instruments. Perhaps this is why existing techniques are not as successful.

Overall, the BayesNet classifier was the best performing classifier and MFCCs were the best performing feature selectors used in the automated based experiments. The use of segmentation outperformed the use of a straight one second sample from which to extract features. This is expected to be due to the alignment of attack portions of notes across all samples. We found that a two-tiered classifier was more successful than a flat classifier.

In the future, as this research has investigated digital based hardware and software musical instruments, questions such as the following could be asked and researched. “Can we classify and distinguish instruments that were created from samples, use FM synthesis, additive synthesis or use any other method of storing digital musical instruments?”. Further exploration of polyphonic based music with respect to unsegmented and segmented audio files, could be undertaken.

7. CONCLUSION

Our research explores techniques for identifying digital musical instruments in audio samples. Through our experiments, we found, that, using multiple short samples after segmentation on note boundaries is far superior to using a continuous sample when classifying instruments in an automatic way. The use of a hierarchical classifier structure gave slightly better results than a flat structure. As in our earlier work, we found that distinguishing pianos was difficult for classifiers. When we tested human perception of instrument timbres, we found that participants had similar difficulties. One hypothesis about the difficulty in distinguishing pianos in our collection, is, that, the variation in timbre within a single instrument is far greater than between two instruments in this category. There may be similar problems with the woodwind and solostring instrument categories, which were also difficult to distinguish for humans. Currently we have no method for measuring timbre variation between instruments, so it is difficult to quantify this across the dataset.

Our results suggest that truly synthetic instruments, such as lead synthesiser sounds, pose difficulties for automatic classifiers, whilst humans have little difficulty distinguishing them. We hypothesise that different features may be required than those used for acoustic instruments, or sounds that closely mimic acoustic instruments.

As one of our objectives is to build “query by timbre” interfaces for digitised music collections, we need to consider human sensitivity to timbre in their design. The results of the experiments discussed in this paper will provide a basis for future work in the design timbre-based retrieval techniques.

8. REFERENCES

[1] G. Agostini, M. Longari and E. Pollastri. Musical instrument timbres classification with spectral features. In *Journal on Applied Signal Processing*, Volume 2003, pages 5–14, 2003.

[2] S. Essid, G. Richard and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. In *IEEE Transactions on Audio, Speech, and Language Processing*, Volume 14, pages 68–80, January 2006.

[3] D. McEnnis, C. McKay, I. Fujinaga and P. Depalle. jaudio: A feature extraction library. In *Proc. of the 6th International Conference on Music Information Retrieval*, pages 600–603, London, UK, Sept 2005.

[4] V. Sandvold, F.Gouyon and P. Herrera. Drum sound classification in polyphonic audio recordings using localized sound models. In *Proceedings of Fifth International Conference on Music Information Retrieval*, pages 537–540, Barcelona, January 2004.

[5] C. Simmermacher, D. Deng and S. Cranefield. Feature analysis and classification of classical musical instruments: An empirical study. In *Proc. of ICDM 2006*, pages 444–458, Leipzip, Germany, 2006.

[6] P. Somerville and A. Uitdenbogerd. Classification of music based on musical instrument timbre. In *Proc 4th Australasian Data Mining Conference*, pages 173–188, Sydney, 2005.