
Ian Kaminskyj

Electrical and Computer Systems

Engineering

Monash University

Clayton, 3800

Australia

ian.kaminskyj@eng.monash.edu.au

Isolated Monophonic Musical Instrument Sound Classification WWW Site

Abstract

The WWW site described in this paper classifies isolated monophonic musical instrument sounds using eight features: cepstral coefficients, constant Q transform frequency spectrum, Morlet wavelets, multidimensional scaling analysis trajectories, RMS amplitude envelope, spectral centroid, vibrato and wavelet packets. Sounds from nineteen instruments of definite pitch, covering the note range C3-C6 and representing the major musical instrument families and subfamilies can be recognized by the system with varying degrees of accuracy and reliability. In addition to identifying instrument sounds, the WWW site also displays 2D waveforms of seven features and 3D waveforms of four features. Java and Java 3D were used to develop a WWW site that is platform independent and provides (i) interactivity, (ii) a graphical user interface and (iii) ease of use. It is hoped the WWW site will assist researchers in both classifying and analyzing the characteristics of musical instrument sounds as well as in music information retrieval.

Introduction

It is now quite common for researchers to develop useful and innovative research tools that they make available to other researchers, either to use on-line or off-line, to help further their work.

Concerning off-line tools, two examples include Marsyas and the Weca system. Developed by George Tzanetakis, Marsyas is a software framework (opihi.cs.uvic.ca/marsyas/) for rapid prototyping and experimentation with audio analysis and synthesis with specific emphasis to music signals and music information retrieval (MIR). Its basic goal is to provide a general, extensible and flexible architecture that allows easy experimentation with algorithms and provides fast performance that is useful in developing real time audio analysis and synthesis tools. Marsyas forms a software framework for developing computer audition algorithms and applications i.e analyze and extract information from audio signals. It provides a general architecture for connecting audio, sound files, signal processing blocks and machine learning. A variety of existing building blocks that form the basis of the most published algorithms in computer audition are already available as part of the framework and extending the framework with new components is possible.

Weca 3 is software developed by Witten and Frank (www.cs.waikato.ac.nz/~ml/index.html) in java that is a

collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from the user's own java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes and is open source software issued under the GNU General Public License. The Weka knowledge explorer forms an easy to use graphical user interface where each of the major packages: filters, classifiers, clusterers, associations, and attribute selection is represented along with a visualization tool that allows datasets and the predictions of classifiers and clusterers to be visualized in two dimensions.

In regard to on-line tools, in the area of analysis and recognition of musical instrument sounds, Studio on-Line (SOL)(forumnet.ircam.fr/402.html?L=1) is a service available from the Institute for music/acoustic research and coordination (Ircam) which provides access to a database of over 21 000 instrumental sounds. Sixteen classical instruments have been recorded note-by-note, over their entire range, with several nuances and by using a setup of four different microphones. The sounds samples have been recorded at 24 bits/48 kHz and are downloadable in aiff format (44.1 kHz, 48 kHz, etc.; 8/16 or 24 bits) facilitated by conversion software written at Ircam. The sound database is accessed via an innovative content-based search and classification interface, called the Sound Palette. The user is able to upload his/her own audio samples and classify them manually or with the help of automatic classification features. S/he can also share his samples with other system users. Access to SOL is via annual subscription or CD on demand purchase

This paper describes an addition to this collection of available research tools. A WWW site has been created that allows users to extract and analyse a number of 2D and 3D features from musical instrument sounds and use these features to automatically identify the instrument source. It is based on work performed by the author on a six feature musical instrument sounds classification system (Kaminskyj and Czaszejko, 2005) which was recently extended to eight features (Pruysers, Schnapp and Kaminskyj, 2005). The corresponding six feature musical instrument sound classification system WWW site (Williams and Kaminskyj, 2002) has now also been extended to cover all eight features. It is thereby hoped that researchers working in the area of analysis and recognition of musical instrument sounds and MIR will be able to use this WWW site to further their research and hopefully, even extend it, by adding extra features and

tools; thereby making it even more powerful, flexible and useful.

Data Collection

Sounds from nineteen musical instruments of definite pitch were used for system development and testing. These are the musical instrument sounds that the WWW site can thereby recognize. The sounds were obtained from the McGill university Master samples (MUMS) CDs and comprised both non-vibrato and vibrato recordings. The instruments include: guitar (plucked string), violin, cello and double bass (bowed string), piano (struck string), flute (air reed wind), accordion, clarinet, saxophone (single mechanical reed wind), oboe and bassoon (double mechanical reed wind), organ (air/mechanical reed wind), trumpet, trombone, French horn, and tuba (lip reed wind) and xylophone, glockenspiel and marimba (percussive definite pitch). Both vibrato and non-vibrato recordings were used for cello, flute and violin. The note range used was C3-C6 of the equally tempered musical scale. Not all instruments cover this complete note range, but each has at least some notes falling within it.

Feature Extraction

The eight features extracted from each musical recording classified by the system are described below.

Cepstral Coefficients

The cepstrum is the Fourier transform of the log magnitude spectrum of the musical sound waveform. There exist many different ways of calculating the cepstral coefficients, principally determined by the spectral measure used. The constant Q transform (CQT) frequency spectrum (Brown, 1991) with quartertone spaced spectral bins was used to calculate the cepstral coefficients (CC_n):

$$CC_n = \sum_{k=0}^{176} \log(X[k]) \cos\left[n(k + 0.5) \frac{\pi}{177}\right]$$

where $X[k]$ is the CQT magnitude for spectral bin k and $n = 1..176$ since CC_{177} is always zero.

Although this equation produces 176 cepstral coefficients, it was determined empirically that for instrument classification purposes, using only the first eleven coefficients produced the best classification results.

CQT frequency spectrum

The CQT frequency spectrum has logarithmically spaced spectral bins aligned with the semitone note frequencies of the musical scale. The spectral bins are spaced at quartertone intervals, which results in 24 spectral bins per octave. The contribution of the frequency component $X[k]$ corresponding to spectral bin k is given by

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k, n] x[n] \exp\{-j2\pi Qn / N[k]\}$$

where, $N[k]$ represents the window length, $W[k, n]$ represents the Hanning window, $x[n]$ represents the n th raw recording sample, and Q represents the quality factor of each 1/24 octave filter bank filter.

The CQT spectrum comprises 177 spectral bins, beginning from C3 (130.81 Hz) and finishing at E10 (21.1 kHz; Nyquist frequency = 22.05 kHz). The frequency corresponding to spectral bin k is given by

$$f_k = 130.81x2^{k/24}$$

As the frequency spectrum of musical instrument tones changes over the duration of a recording, in this research, the spectrum is measured at the point in time where the sum of the spectral bin amplitudes reaches a maximum, thereby where the maximum signal to noise ratio is reached.

Morlet Wavelets

Morlet wavelet analysis (MWA) was performed using Morlet wavelets because of their optimal time-frequency resolution (Pruysers, Schnapp and Kaminskyj, 2005). MWA was performed using the continuous wavelet transform, which provided logarithmic spacing of wavelet band pass filter (BPF) centre frequencies. Wavelet BPFs were centred at frequencies corresponding to each note of the equally tempered scales, i.e. 12 filters per octave. Each wavelet coefficient thereby corresponded to a specific note and is calculated using

$$z_{i,j} = \int_{-\infty}^{\infty} x(n) \cdot 2^{i/12} \cdot y_i(n-j)$$

where $z_{i,j}$ relates to the wavelet coefficient at frequency i and time interval j , $x(n)$ represents the raw recording samples and $y_i(n-j)$ relates to the Morlet wavelet at frequency i and time interval j .

A BPF quality factor of 51.5 was chosen as a compromise between having filters with too large a bandwidth and lacking good frequency resolution, and filters with too small a bandwidth, missing harmonics of incorrectly tuned tones. Extracted wavelet features were also frequency normalized by analyzing 3.5 octaves of spectral data, starting at the fundamental frequency. In order to reduce data dimensionality and remove phase complications, wavelet coefficients were calculated (a) using only the first 0.75 sec of a recording, (b) following root mean square (RMS) averaging, (c) subsampling the resultant wavelet coefficients and (d) only at harmonic frequencies.

MSA trajectories

The multidimensional scaling analysis (MSA) trajectories used in this research are based on work performed by Hourdin et al. (1997) who applied MSA techniques to physical descriptions of musical sounds to study timbre. They have, however, two important differences to the approach taken by Hourdin et al. Firstly, CQT analysis

was used to obtain a musical instrument tone spectral representation. Secondly, principal component analysis (PCA) (Jolliffe, 1986) was applied instead of factorial analysis of correspondences.

To obtain the desired instrument trajectories, MSA was performed on 3-dimensional CQT waterfall plots, comprising 48 CQTs performed sequentially on each 1 second recording using a step size of 15.625 msec. Extracts were then taken that included only the spectral bin of the first 20 harmonics as well as a spectral bin on either side. This approach reduced the number of spectral bins analysed for any given spectral snapshot from 177 down to 53.

Upon applying PCA to the 3-dimensional CQT waterfall plot extracts for all musical instrument sounds allowed the first three principal components to be used to define a 3-dimensional space in which musical instrument trajectories could be displayed over the whole one second duration of a tone. Empirical results showed that best classification results were obtained using amplitude normalised trajectories (the maximum (x,y,z) coordinate absolute amplitude being normalised to one).

RMS amplitude envelope

The formula used to calculate the RMS amplitude envelope is:

$$A(n) = \frac{1}{A_{\max}} \sqrt{\frac{1}{N} \sum_{i=1+nN}^{(n+1)N} x^2(i)}$$

where, $A(n)$ represents the RMS amplitude envelope, A_{\max} is the maximum amplitude of the amplitude envelope, $x(i)$ represents the raw recording samples, and N is the window size over which each RMS calculation is performed.

The most appropriate value used for N was empirically found to be $N = 3Tp$, where Tp represents the period of the waveform fundamental calculated as a number of samples. Since the pitch of each input monophonic recording is known *a priori*, N can be easily calculated before evaluating $A(n)$.

Spectral Centroid

The spectral centroid, also called brightness, is essentially a measure of the 1st moment of the spectral energy distribution. It is calculated over the duration of musical instrument recording using:

$$S_k = \frac{\sum_{j=0}^{176} A_{j,k} f_j}{\sum_{j=0}^{176} A_{j,k}}$$

where S_k is the spectral centroid at time interval k , $A_{j,k}$ is the amplitude of CQT spectral bin j at time interval k , and f_j is the frequency of CQT spectral bin j .

Vibrato Detection

The vibrato detector uses the following conditions to qualify a tone as having been played with vibrato:

- the tone was produced by a sustain instrument,
- the peak-to-peak amplitude of the frequency variation of the largest amplitude harmonic exceeded 20 cents, and
- the frequency modulation of the tremolo waveform or the vibrato waveform of the largest amplitude harmonic following the harmonic amplitude tracking scheme and post processing compensation algorithm (Kaminskyj, 2005) occurred at a rate of 5-8 Hz (as measured by the location of the peak of the frequency spectrum).

These conditions were based on the assumption that only cello, flute and violin were capable of being played with vibrato.

Wavelet Packets

Wavelet packet analysis (WPA) decomposes a signal into “packets” by simultaneously passing the signal through a low decomposition filter (LDF) and a high decomposition filter (HDF) in a sequential tree like structure (Pruysers, Schnapp and Kaminskyj, 2005). Of the large number of filters types that can be used for this purpose, three were empirically chosen and evaluated to see which provided the best classification results. For non-vibrato recording classification, Daubencies 20 wavelet packets proved optimal, while for vibrato recording classification, Coifman 5 wavelet packets were the best. The output of each filter is described by the convolution process

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k)$$

where $y(n)$ is the filter output, $x(n-k)$ represents the raw recording samples, $h(k)$ relates to the filter coefficients and N indicates the number of filter coefficients.

Passing a signal through a pair of LDF and HDF filters produces two packets: (1) the Approximation and (2) the Detail respectively. This is referred to as level 1 decomposition. The level 1 packets can then be passed through another pair of filters to produce a total of 4 packets (level 2 decomposition). This operation can be continued indefinitely, although after a certain point, which has to do with the signal sample length, the packets from one instrument become less distinguishable from that of other instruments, affecting classification accuracy. It was found empirically that the optimal number of levels to use for classification purposes was between 3 and 5. Therefore, this meant that between 2^3 and 2^5 packets were extracted for music tones comprising 88200 samples. As for MWA, in order to reduce data dimensionality, wavelet coefficients were calculated (a) using only the first 0.75 sec of recording, (b) following RMS averaging and (c) subsampling the resultant wavelet coefficients.

Classification

The classifier structure is shown in Figure 1.

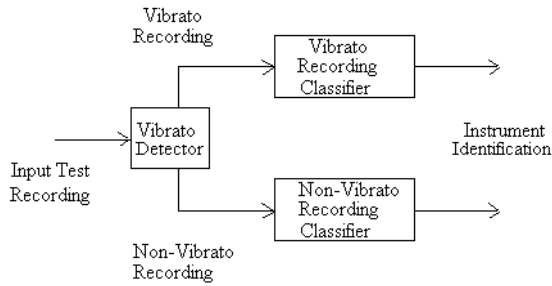


Figure 1. Classification system structure

The vibrato detector firstly determines the presence of any vibrato within the input test recording. If vibrato is deemed to be present, the vibrato recording classifier is invoked, which classifies the input as having been produced by one of three possible vibrato instruments: cello, flute and violin. Alternatively, if no vibrato exists, the non-vibrato recording classifier is executed. It classifies the input as having been produced by one of nineteen possible non-vibrato instruments.

Given the increased complexity of the non-vibrato recording classifier, in terms of the number of instruments it needs to classify, three different classifier architectures were evaluated: (a) single stage, (b) hierarchic, and (c) hybrid. It is intended that eventually, the user of the system will be able to decide which s/he wishes to use, balancing the performance achieved with the computational effort expended. For the purposes of the current WWW site implementation, only the single stage classifier has been implemented.

The single stage classifier shown in

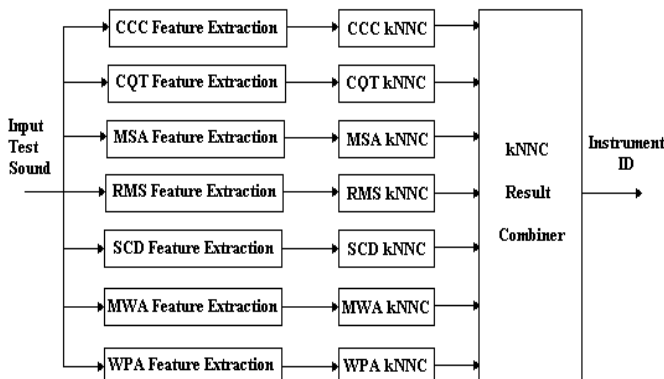


Figure 2. Single Stage Classifier

Figure 2 uses seven single feature classifiers to independently determine, based on their single feature, the most likely instrument to have produced the input test sound. Each classifier extracts its single feature and then employs the k NNC algorithm (Duda and Hart, 1973) to classify the input test sound.

The single features were used either directly or pre-processed using PCA. The k NNC result combiner uses a confusion matrix for each of the single feature classifiers to summarise their strengths and weaknesses in terms of classification accuracy and reliability. The combination process aims to achieve the best possible overall result when combining the individual classifier results. Finally, search limits and instrument elimination techniques were introduced in an attempt to improve the classifier performance.

WWW Site Graphical User Interface

Refer to Figure 3 which summarises the layout of WWW site graphical user interface (GUI) and what each section is used for.

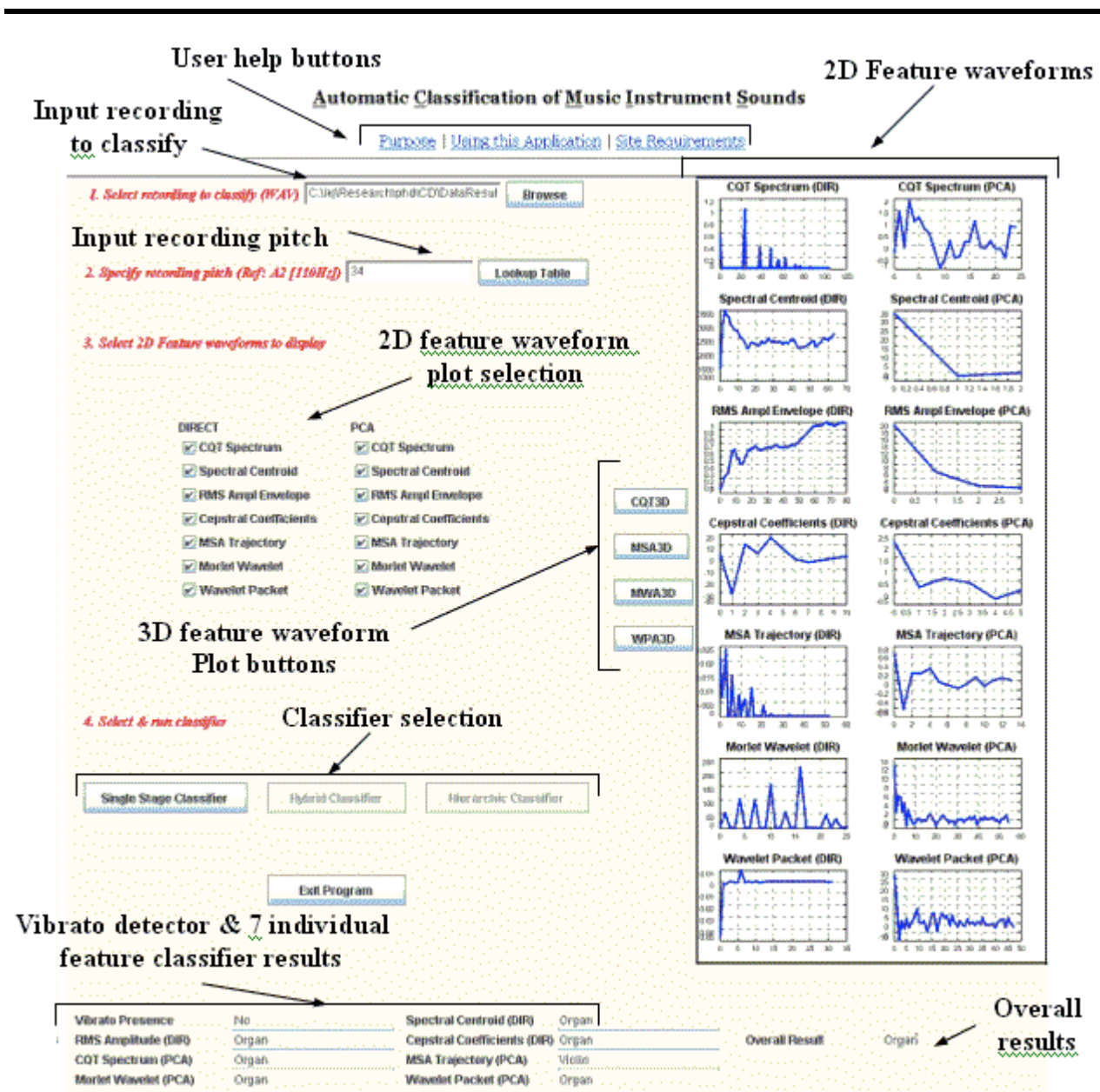


Figure 3. Graphical user interface layout and functionality

This WWW site can be accessed from the author's home page at users.monash.edu.au/~kaminski/.

Sample Session

To help clarify how the WWW site would typically be used, a hypothetical sample session will now be described, where a user would like to classify a recording of a pipe organ playing note G5, obtained from the MUMS CD collection.

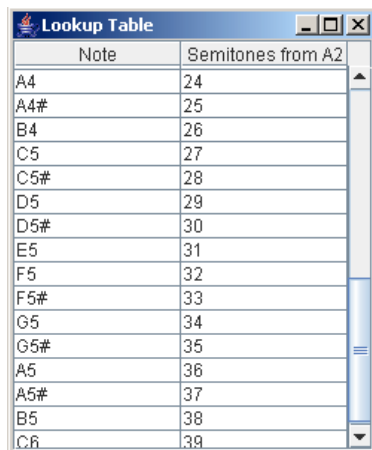
Referring to Figure 3, if the user were unfamiliar with the WWW site, s/he may begin by accessing the on-line help via the buttons **Purpose**, **Using the Application** and **Site Requirements**. The help pages that pop-up in response to these button activations indicate respectively: (i) the purpose of the WWW site, (ii) how the user would typically go about using the WWW site, and (iii) what are the site requirements, in terms of

browser plug-ins required to allow the user's browser to run the WWW site applet. Once familiar with the WWW site, this information would then only rarely be accessed by the user.

Referring to Figure 3, the user would then begin by specifying the recording to be classified in the *1. Select recording to classify (WAV)* textbox. If required, the user is able to browse their computer hard drive to find the location and name of the file containing the recording, simply by activating the **Browse** button.

With the recording selected, the user then needs to specify its pitch, in terms of the number of semitones it is above note A2. For our organ tone at G5, this would correspond to a numerical semitone difference of 34. This value needs to be entered into the *2. Specify recording pitch (Ref: A2 [110 Hz])* textbox. If necessary, the user can also determine this value by activating the **Lookup Table** button to obtain the Table shown in Figure 4 and

scrolling down until the numerical pitch value associated with the G5 note is shown. As the classification software requires the recording pitch information to perform the classification task, currently the user needs to supply this information manually.



Note	Semitones from A2
A4	24
A4#	25
B4	26
C5	27
C5#	28
D5	29
D5#	30
E5	31
F5	32
F5#	33
G5	34
G5#	35
A5	36
A5#	37
B5	38
C6	39

Figure 4. Look up Table for entering recording pitch information

The next task for the user is to determine which 2D feature waveforms to display, via the 3. *Select 2D Feature waveforms to display* checkboxes. As features are used both directly as well as after PCA for classification purposes, the user is able to either look at both of these for all features or only those of interest. In the example above, the user has elected to display both the direct and PCA 2D feature waveforms for all seven features.

Finally, the user needs to select which classifier to run to classify the input recording. Currently, only the single stage classifier is supported, so the user can only activate it by activating the **Single Stage Classifier** button. It is hoped that in the next revision of the WWW site, all three classifiers will be supported.

After a few minutes, the result of the classification process is displayed. As can be seen on the right hand side of the GUI, all the 2D feature waveforms that the user selected are displayed. If the user wishes to also examine any of the 3D feature waveforms, this can be done simply by activating any of the four 3D feature waveform buttons (**CQT3D** for CQT frequency spectrum, **MSA3D** for MSA trajectories, **MWA3D** for Morlet wavelet analysis or **WPA3D** for wavelet packet analysis). Figure 5 shows all four 3D waveforms for the example pipe organ G5 tone, which are displayed in pop-up windows on the screen for the user to manipulate (eg. resize, move etc), totally independently of the main GUI page.

At the bottom of the screen, the results of the vibrato detector are shown together with those of the seven individual feature classifiers (refer to

Figure 3). For this example, the vibrato detector determined that the pipe organ G5 tone was not played with any vibrato. Looking at the individual feature classifier results, it can be seen that of the seven classifiers,

only one did not classify the input recording as an organ, the MSA trajectory feature classifier. On this basis, the result combiner had no difficulty ascertaining that the input recording was produced by a pipe organ. At this point, the user would be free to perform another classification if required, or leave the WWW site.

Conclusion

This paper has described a WWW site that has been created to allow users to extract and analyse a number of 2D and 3D features from musical instrument sounds and use these features to automatically identify the instrument source. Much work still remains. Future work will include (a) implementing the hierarchic and hybrid classifiers, (b) providing the user with the ability of modifying the classifier parameter settings, to investigate how these parameters influence classifier performance, (c) provide an automatic pitch detection facility, and (d) support different sound file formats. It is thereby hoped that researchers working in the area of analysis and recognition of musical instrument sounds as well as MIR will be able to use this WWW site to not only further their research but hopefully, to extend it by adding extra features and tools; so as to make it even more powerful, flexible and useful.

Acknowledgements

I would like to acknowledge the contributions of the following final year project students who, under my supervision, have contributed to the WWW site development over recent years: Mukesh Muruganandan, Kee Hon Chan, Serdar Yilman, Kang Chan, Ho Pang Sie, Dawei Yin, Thanh Tran and Mark Williams.

References

- Brown, J. C. 1991. "Calculation of a constant Q spectral transform" *Journal of the Acoustical Society of America*. 89. 425-34.
- Duda, R. O. and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. Wiley-Interscience.
- Hourdin, C. Charbonneau, G. et al. 1997. "A multidimensional scaling analysis of musical instruments' time-varying spectra" *Computer Music Journal*. 21, 2. 40-55.
- Jolliffe, I. T. 1986. *Principal Component Analysis*. Springer-Verlag,
- Kaminskyj, I. 2005. "Automatic Recognition of musical instruments using isolated monophonic sounds" *PhD thesis*, Monash University.
- Kaminskyj, I. and Czaszejko, T. 2005. "Automatic Recognition of Isolated Monophonic Musical Instrument Sounds using kNNC" *Journal of Intelligent Information Systems, Special Issue on Multimedia Applns.* 24, 2-3. 199-221.
- Pruyters, C, Schnapp, J. and Kaminskyj, I. 2005. "Wavelet Analysis In Musical Instrument Sound Classification" *8th International Symposium on Signal Processing and its Applications*. Aug. *Studio online Project*. forumnet.ircam.fr/402.html?L=1 (6th December 2005).

Tzanetakis, G. *Marsyas software*.

opihi.cs.uvic.ca/marsyas/ (6th December 2005).

Weca project.

www.cs.waikato.ac.nz/~ml/index.html (6th December 2005).

Williams, M. and Kaminskyj, I. 2002. "WEB Based Automatic Classification of Musical Instrument Sounds" *Proc. Aust. Comp. Music Assoc. Conf.* Melbourne, Australia.

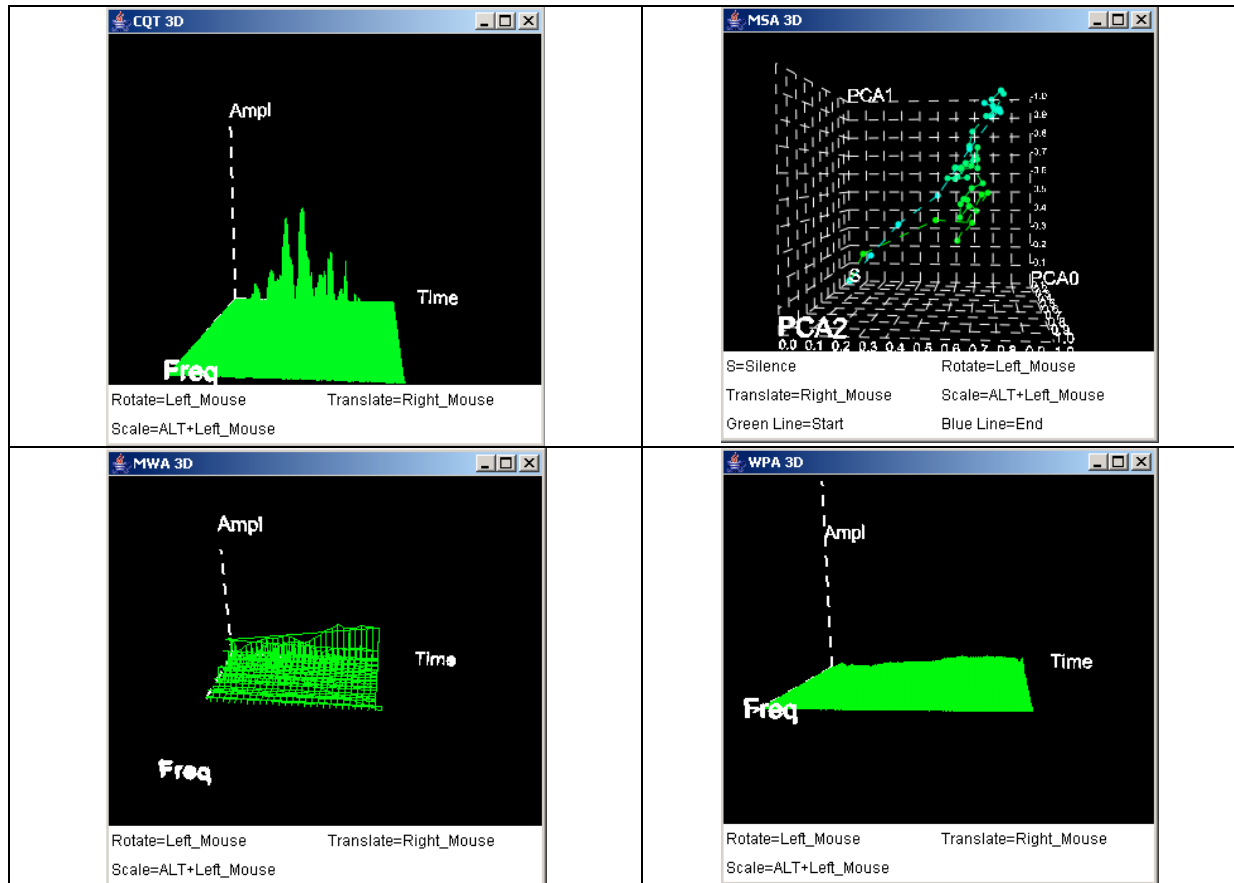


Figure 5. 3D Feature waveforms for pipe organ tone, pitch G5