Violin Note Classification

J. A. Charles, *Student Member, IEEE,* D. Fitzgerald, E. Coyle

***Abstract* — Although much research has been carried out on finding features for instrument recognition systems, little work has focused on specifically the violin's entire timbre space. Suitable features from which a computer can assess the quality of a violinist's playing have been sought and the classification of violin note sound quality is investigated in this paper. The eventual outcome of this work can be applied in various systems including the development of a violin or bowed string instrument teaching aid, in automatic music transcription and information retrieval or classification systems.**

***Keywords* — classification, clustering, data analysis, data representation, perception, sound analysis, violin timbre.**

I. Introduction

To gain better understanding of the relationship between violin playing technique and the sound produced, a suitable means of quantifying and classifying these differences is needed. This is so that guidelines may be established for not only good violin sound but also for poorer or beginner violin playing with the ultimate aim of developing a computer based violin teaching aid, of which none exists. The more general area of quantifying beginner and professionals standard legato violin note samples using signal processing techniques was presented in [1, 2]. This has enabled the representation of violin sounds by suitable descriptors. Violin playing faults have been identified and are limited to nine faults at this stage. This paper considers the classification of violin notes using up to fifteen features. Two tasks are put to a k-means nearest neighbour classifier: the first is the detection of beginner note from a professional standard note and the second is much more specific, involving individual fault detection.

In the following sections, existing research is briefly presented, followed by a description of the data set requirements and how it was obtained, after which the listening tests are detailed. The choice and brief explanation of the features used in the classifier are then given, followed by the classification method and results obtained.

II. Existing Research

Much existing research on violins has been carried out in order to better understand and emulate the making of top quality sounding instruments. Many methods have been applied to gain insight into the complex interactions between the various components of stringed instruments. Work is ongoing considering the problem of quantifying perception relating to violin sound quality [3]. However, work exploring the effect a player has on the violin sound produced is limited. Finding features which are suitable for quantifying the violin's timbre space involves exploring the effect of a player on sound quality. Many features, although very useful in determining one instrument from another [4, 5], are not appropriate for representing the subtleties due to playing technique or for use within an individual instrument's timbre space.

III. Data Set

As no suitable data set was readily available, one had to be made. Much thought was given in creating this data set in terms of what was needed, obtainable and viable. The ideal data set would be a type of violin timbre real sound continuum. Unfortunately, this would be very time consuming, if not near impossible to obtain. The first bow stroke a beginner must learn is *legato*, which literally means 'tied together' or smoothly connected [6] as opposed to slurred, which refers to multiple notes in a same bow stroke. Although legato playing encompasses all lengths of bow stroke, in this work it means using full bows. Mastering this ensures enough bow control upon which the student can develop other bow strokes, such as *staccato* ('disconnected' [6]). Since the style or type of bow stroke used effects the readings obtained, only professional standard player *legato* notes will be used and the beginner notes will be compared to these.

The data test set consists of two same sized groups, one with beginner notes and the other with professional standard player legato notes. The samples all contain one note and are of varying lengths and pitches, making it more appropriate to use features which do not dependent on ether note length or pitch. The data samples were obtained in a recording studio using a stereo pair of dynamic microphones, a condenser microphone switched to omni, and a large diaphragm condenser microphone also with omni-directional pick-up pattern. The tracks were recorded onto DAT, mixed and saved as monophonic wav files. The recordings were all made in the same studio, using the same microphones and set up as well as the same violin and bow.

IV. Listening Tests

Listening tests have been included to remove the subjective nature of this research by showing that other trained string players can hear and recognize the faults and

Financial support was received for this research in part from Dublin Institute of Technology.

J. A. Charles is a graduate student at the School of Electrical Engineering Systems, Dublin Institute of Technology, Kevin St., Dublin 8, Ireland (phone: +353-86-8338033, email : violincharles@gmail.com)

sound quality descriptions used and most importantly, that a priori labels for the classifier may be obtained. From the results of these listening tests, it is hoped that a relationship can be established between what people perceive and any quantitative features for the sound samples. These tests are aimed at professional standard violinists in particular but, to increase numbers, cellists and violists have also been included.

The listening group consisted of twenty-one string players. The listeners received no training, only a copy of the testing process steps and an explanation of the terms. A play list which includes all the beginner and legato good note samples, 176 samples in total, exists. As soon as the listener activates the testing/listening program, a random play list is generated consisting of all samples from the list. After having heard the note, the listener selects the terms which best characterise the sound and grades the overall quality. The sound characteristics list includes descriptions of playing faults and the overall sound quality is a grade between 1 (very poor) and 6 (excellent). The faults or sound characteristics are crunching, skating, nervousness (uncommitted, faltering sound), intonation, bow bouncing, extra note, sudden end to note, poor start to note and poor finish to note.

The exact play list for each listener only becomes available at the end of the listening test. The test progresses at a speed controlled by the user and each sample can only be played once. AKG K240 ‘Monitor’ (600 Ohms) headphones were used and samples were accessed and played through Matlab. The consistency of the results obtained from this test were checked and found to be acceptable. Normalising these results allowed for an average listener to be established. This average listener is what is used for investigating how violin timbre is perceived and for use as the a priori sample labeling in the classifier.

V. Features

The efficacy of many features from the time, spectral and cepstral domains have been tried for their ability at representing change within the violin timbre. Based on the visual inspection of these results and their ability at differentiating between beginner note and professional standard legato notes, features were selected for use in the classifier. The fifteen features selected are given in Table 1.

Table 1: Features Used.

Feature	*Description*
1	Time Domain Mean (TM)
2	Time Domain Kurtosis (TK)
3	CQT Harmonic Strength (CQTH)
4	PSD <190Hz (PSD190)
5	Spectral Flatness Measure Mean (SFMM)
6	Spectral Flatness Measure Variance (SFMV)
7	Spectral Contrast Measure <190Hz (SCM190)
8	Real Cepstral Coefficients Mean (RCCM)
9	Real Cepstral Coefficients Variance (RCCV)
10	Real Cepstral Coefficients Kurtosis (RCCK)
11	1st Real Cepstral Coefficient (RC0)
12	2nd Real Cepstral Coefficient (RC1)
13	6th Real Cepstral Coefficient (RC5)
14	Spectral Centroid Variance (SCV)
15	Mel Cepstral First Coefficient Mean (MC0M)

Although up to eighteen feature vectors have been used previously in the classifier as it is, exceeding fifteen features is not practical for time and computing reasons. All of these features group the beginner samples in a visually discernable way from the professional standard legato notes, which was certainly not the case for most of the features tested. These features are all standard but some have been modified slightly to better suit the research aims, such as the PSD and spectral contrast measures below 190Hz, the first of which can be seen in Figure 1. The lowest note on a violin tuned to A440, is the open G string at ≈196Hz. Looking at the frequency content below the violin’s frequency range was done in the expectation that information relating to the lack of playing ‘neatness’ might be revealed, as can be seen in Figure 1 where the beginner samples have more frequency content below the lowest note than the professional standard notes do.

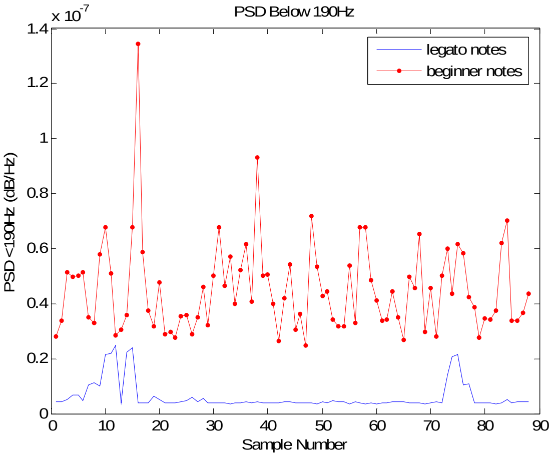

Fig. 1. Spectrum Power Present Below 190Hz.

Although the features given in Table 1 split the data set into distinct groups, three features completely separate the sample groups. They are the time domain mean, the CQT harmonic strength and the spectral contrast measure below 190Hz. How these features perform in a classifier is presented next.

VI. Classification

Classification is the general term given to organizing or grouping similar data together according to selected characteristics or some common feature. Grouping data together based on similar patterns or descriptive features allows a class label to be associated with the group. The most significant aims of classification relate to data simplification and prediction, increasing the efficiency of tasks such as information retrieval [7]. In this paper, the classification of note samples into beginner or professional and fault identification are tested. The aim is to provide objective and stable classification for the subjective nature of violin sounds.

The first stage of the classification process involves clustering which is used to find centres that reflect the distribution of data points [8]. Running the k-means clustering algorithm provides the prototype vectors which are then used in the k-NN classifier. Although many clustering methods exist, k-means is one of the most often used because of its simplicity and converges well with the Euclidean distance which is given in equation 1 [8].

$$dist = \frac{1}{N}\left(\sum_{n=1}^{N}(A(n)-B(n))^2\right)^{1/2} \qquad (1)$$

One advantage of using the Euclidean distance is that each feature remains equally important and no correlations between variables influence the outcome. The k-means clustering code, taken from the Somtoolbox [9], uses the iterative partitional clustering algorithm put forward by Jain and Dubes, a description of which can be found in [8]. An advantage of this algorithm is that it automatically assigns items to clusters. The disadvantages are that the number of clusters must be pre-selected and that all items are forced into a cluster, making it very sensitive to outliers. The squared Euclidean distance metric is used which is computationally faster for clustering than the Euclidean distance shown in equation 1. The clustering algorithm remains unaffected by this change, as it is a partitional clustering method, as opposed to a hierarchical one.

For the first task, two clusters are sought: one for poorer quality sounds and another for professional violinist notes. The ‘beginner’ and the ‘professional’ clusters provide the k-NN classifier with its prototype vectors. They are two 15 x 1 vectors. For the fault identification task, clusters are formed according to the presence or absence of a particular fault as perceived by the listener. Prior to use in the classifier, these cluster vectors were checked by comparing their values with the means of all samples for each feature associated with its respective cluster. The algorithm converged well and no alterations had to be made.

The data set’s features are stored in a 176 by 15 array, where 176 is the total number of samples and 15, the number of features. A proximity matrix is then calculated using the squared Euclidean measure between the prototypes and each feature vector. This matrix is the input for the k-NN classifier, to which class labels are assigned. These labels are then compared with the a priori labels to obtain the classifier accuracy reading. Classifier accuracy is the probability of correctly labeling a randomly selected sample. The k-NN rule classifies a sample by assigning it the label which is most often associated with its k-nearest samples. When k=1, every sample is assigned to the class of the nearest cluster or pattern. In practice, k=1 is often used, as it is in this work.

Should the classification process be carried out on the entire data set, very specific model building information will be obtained. Cross-validation techniques are methods for detecting and preventing classifier over-fitting, checking classifier accuracy estimation and generalisation potential. It is a way of ensuring that a classifier can perform in an unsupervised situation. To conduct cross-validation, the data set is put in a random order after which, a portion of the data set is put aside as a ‘training’ set and leaving the rest for testing. In n-fold cross-validation, the data set in put into n equal sections where n-1 sections are used for training and the remaining section for testing. The means are taken of the n folds. Four-fold cross-validation is used in this work.

VII. Results

All results have been obtained using four-fold cross validation for both tasks and are presented below.

A. *Beginner vs. Professional*

A summary of the results obtained for the detection of beginner from professional standard legato notes based on the number of features used is given in Table 2. From the results returned using the features given in Table 1, it is possible to detect a beginner note from a professional standard legato note with 96.59% accuracy using just one feature. Two features returned this result; they are the time domain mean and the CQT harmonic strength. Using all fifteen features return detection rates of 91.10% accurate.

Table 2: Task I Results Summary.

No. Features	*Test %*	*Train %*
1	96.59	96.59
2	96.59	96.59
3	96.59	96.59
4	96.59	96.59
5	96.59	96.59
6	96.59	96.59
7	95.45	95.45
8	95.45	95.45
9	95.45	95.45
10	95.45	95.45
11	95.45	95.45
12	93.94	92.05
13	93.18	92.05
14	92.61	91.48
15	91.10	90.91

The monothetic results can be seen in Table 3. This has been included so that the performance of each feature can be seen. Of interest in this table are the results returned when features TM, CQTH and SCM190 are used as these features separated the data set 100% accurately in their respective domains, indicating that a simple threshold value could be used.

Table 3: Monothetic Results Task I.

Feature	*Test %*	*Train %*
TM	**96.59**	**96.59**
TK	91.86	91.48
CQTH	**96.59**	**96.59**
PSD190	51.70	50.57
SFMM	59.85	53.41
SFMV	83.33	85.80
SCM190	**92.05**	**92.05**
RCCM	91.48	91.48
RCCV	90.34	90.34
RCCK	86.36	88.64
RC0	87.88	88.07
RC1	90.34	90.34
RC5	87.5	86.36
SCV	67.05	64.77
MC0M	48.48	46.59

When features TM, CQTH and SCM190 are used in the classifier, although the three highest results are returned, they are not 100%. A possible explanation for this is the sensitivity of the Euclidean distance to outliers in the data.

B. *Fault Detection*

Individual fault detection using this feature set did not prove to be very effective. Only one fault, player nervousness, was detected to 73.67% train and 77.84% test. Increasing the number of features in this task did not significantly alter the results but caused the error reading to decrease. Using twelve and thirteen features returned 72.54% and 72.16% respectively on its training set and 71.02% on the test sets. The features returned for detecting player nervousness detected other faults too but a gap of at least 10% exists between its detection and the detection of any other fault.

Playing fault detection was improved by using different features. Using features which did not necessarily group the beginner notes and the legato professional standard notes distinctly, improved the results for fault detection. The features used were: TM, moving mean variance, RCCM, RCCV, RCCK, RC0, RC1, RC3, SCM, SFM, SFMM, SFMV, SFM skew, PSD, autocorrelation function. Using combinations from this feature list, it was possible to detect playing faults as can be seen in Table 4.

Playing faults bow bouncing and extra note, achieved the highest accuracy readings at 90.15% and 90.34% on their respective training sets, 92.61% and 89.77% on their test sets. Crunching, bow bouncing, extra note, poor start and poor end to note all returned the same feature combination associated with their respective top detection readings. Skating, poor intonation and sudden end to the note all use the same ten feature combination to achieve detection. Although detection results are lower for nervousness, the feature combinations do not overlap at detecting any of the other faults, making it also possible to detect nervousness.

TABLE 4: TASK II RESULTS SUMMARY.

No. Features	*Train %*	*Test %*	*Fault Detected*
1	73.48	76.14	nervousness
2	83.33	86.36	bow bounce
3	87.12	85.80	bow bounce
3	87.69	89.77	extra note
4	90.15	92.61	bow bounce
4	90.34	89.77	extra note
5	89.77	90.34	bow bounce
5	90.34	89.77	extra note
6	87.12	85.80	bow bounce
6	87.69	89.77	extra note
7	87.12	85.80	bow bounce
7	87.69	89.77	extra note
8	87.12	85.80	bow bounce
8	87.69	89.77	extra note
9	83.33	86.36	bow bounce
10	83.14	85.80	bow bounce
11	83.33	86.36	bow bounce
12	83.33	86.36	bow bounce
13	83.33	86.36	bow bounce
14	83.14	86.36	bow bounce
15	60.23	60.23	nervousness

Results of above 83% are returned when two to fourteen features are used, all detecting bow bouncing and extra note. Fault detection is less conclusive when fifteen features are used as all training and testing sets return about 60% accuracy. The results obtained for detecting bow bouncing and extra note are very close as can be seen in Table 4. The results for detecting extra note are always marginally higher than those returned for detecting bow bouncing when two to ten features are used. However, using ten to fourteen feature combinations provides solutions for detecting bow bouncing only. All ten, eleven and twelve feature combinations provide results that are at least ≈8% higher for detecting bow bouncing than for the detection of any other fault. The thirteen and fourteen feature combinations return results for detecting bow bouncing which are at least ≈5% higher than for any other fault. Bow bouncing can be detected to above 83% accuracy.

VIII. CONCLUSION

Detecting good sound from beginner sound can be achieved returning accuracy results of just below 97% using one to six features taken from the feature list given in Table 1. Much feature redundancy is present though as all combinations have either feature one or three present, both of which return 96.59% accuracy on their own.

The presence of playing faults can be detected but individual faults are harder to isolate. Only three specific faults can be detected independently so far. They are nervousness, using the feature list given in Table 1, bow bouncing, and extra note which use a different feature list, given in Section VII.B. The detection accuracy rates for the other faults are all closely grouped together, and return the same feature combinations. This is due in part to a sonic similarity between certain faults and that the samples in the data set often contain more than one fault. One possible way around this would be to use samples which contain only one fault at a time but this will be difficult as playing faults rarely occur independently. Another would be to find new features. Location dependent features, such as those pertaining to the attack and end of note periods could be more informative.

The results for both tasks have been obtained via cross validation on one data set. It remains to be confirmed whether they hold on a different data set.

ACKNOWLEDGMENT

Special thanks go to Lioba Petrie, Ailleen Kelleher, Kitty Dalligan, Erica Davidson, Grainne Hope, Sinead Hope, Karla Charles and Owen Tighe.

REFERENCES

[1] Charles, J. A., *et al.* ‘Quantifying Violin Timbre’, DMRN, 2006.

[2] Charles, J. A., *et al.* ‘Violin Timbre Space Features’, ISSC’06, Dublin Institute of Technology, Dublin, June 28-30, 2006.

[3] Woodhouse, J., Cross, I., Moore, B.C.J., Fritz, C. ‘Perceptual Tests With Virtual Violins’, Proc. Institute of Acoustics, Vol. 28, Pt. 1, 2006.

[4] Eronen, A., Klapuri, ‘Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features’, Signal Processing Lab, Tampere.

[5]Martin, K. D., Kim, Y. E., ‘Musical Instrument Identification: A Pattern-Recognition Approach’, 136th Meeting ASA, Oct. 1998.

[6] Jackson, B. G., Berman, J., Sarch, K. *The ASTA Dictionary of Bowing Terms for Stringed Instruments*, American String Teachers Association, 3rd edition, Tichenor Publishing Group, Bloomington, 1987.

[7] Bishop, C. M., *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1996.

[8] Jain, A. K., Dubes, R. C., *Algorithms for Clustering Data*, Prentice-Hall, 1988.

[9] www.cis.hut.fi/projects/somtoolbox/