Violin Timbre Space Features

J. A. Charles[ϕ], D. Fitzgerald[*], E. Coyle[ϕ]

[ϕ]School of Control Systems and Electrical Engineering,
Dublin Institute of Technology,
IRELAND
E-mail:[ϕ]jane.charles@dit.ie
Eugene.Coyle@dit.ie

** Department of Electronic and Computer Engineering,*
Cork Institute of Technology, Cork
IRELAND
*E-mail: *derry.fitzgerald@cit.ie*

Signal processing techniques, from which the quality of a violinist's playing can be assessed by a computer, are presented in this paper.

Keywords – violin, CQT, harmonics, cepstral analysis, spectral centroid, spectral flux, PSD, SFM.

I INTRODUCTION

Steps towards the development of a violin teaching aid which is based on violin pedagogy, sound analysis, and comparison of beginner and good player recordings was presented in [1]. The relationship between timbre and playing technique has been explored and five main beginner faults have been determined. Briefly, the tone fault categories are onsets, offsets, amplitude, unevenness and asymmetry about the x-axis which may contain undesirable sounds such as squeaks, crunches, skating and nervousness. Features which best describe these faults for classification purposes are considered in this paper. This involves getting a suitable set of features which can describe quantitatively the qualitative and subjective nature of violin playing quality. Many features, although very useful in determining one instrument from another [2, 3], are not appropriate for catching the subtleties due to playing technique or for use within a timbre space. Results have been obtained clearly showing that it is possible for a computer to differentiate between recordings of a beginner note and a good player legato note played on a violin [4]. Further signal processing methods will be considered in this paper to find features which best describe violin sound within its timbre space.

II EXISTING RESEARCH

Current advances in signal processing and interactive computing have enabled the development of much more sophisticated systems and learning aids. Hämäläinen *et al.* developed a successful real-time singing aid in [5], which describes the use of pitch-based control of a game character by the user's voice. However a direct transfer of this approach into a violin, or another instrument teaching aid wouldn't be as successful. A singer is physically 'free' to concentrate on a screen and able to react to it. Instrumentalists, especially beginners, need to be looking at what they are doing and looking elsewhere, such as at a screen, will disturb their position. For this reason, a system which offers feedback after the user has played their short piece would be much more effective. This differs greatly in approach to the Music Minus One [6] CDs which offer a variety of recordings to which the user plays the solo part. There seems to be no work conducted on poor violin technique, its effect on sound or on the more general area of the violin timbre space affected by a player using signal processing techniques. Some, but not much work has been conducted on poor singing with the information retrieval domain [7, 8].

III DATA TEST SET

The data test set consists of two same sized groups, one with beginner notes and the other with good player legato notes. The files all contain one note and are of varying lengths and pitches. There are eighty-eight beginner note files and eighty-eight legato good note files. A player will never play two notes exactly the same although they may be perceived by a listener as being the same. A beginner does not have the control necessary to achieve this level of accuracy in playing. Hence, it is more appropriate to not dependent on ether note length or pitch. The ultimate aim is to find features for fault detection within the violin timbre space, which can be applied to the note independent of its length or

pitch. The data files were made in a recording studio using two microphones, one directional, the other, omni directional. The tracks were recorded onto DAT, mixed and saved as monophonic wav files. It should also be noted that 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 VIOLIN TECHNIQUE AND SOUND

The first bow stroke a beginner must learn is called *legato*, which literally means 'tied together' or smoothly connected [7]. Mastering this ensures enough bow control upon which the student can develop other bow strokes, such as *staccato*. Initially the aim would be based on developing a student's legato bow stroke. Since the style or type of bow stroke used effects the readings obtained, only good player *legato* notes will be used and the beginner notes will be compared to these.

V FEATURE EXTRACTION

Features can be considered as descriptors and standard features for extracting information pertaining to musical signals include pitch, spectral centroid, zero-crossing rates, mean acoustic energy, onset, offset times to name but a few. In [3], many features have been determined. Many features, although very useful in determining one instrument from another, are not appropriate for understanding the discrepancies due to playing technique within an instrument's timbre space. Pitch related or dependent features are of limited use within the context of bowing. Through visual inspection of the good player waveforms compared to ones produced by the beginner player, the latter files were much more asymmetric. No real violin sound produces perfectly symmetric waveforms. This is due to the physics of the instrument and the large number of variables which effect the waveform. This asymmetry led to investigating skew readings for these files. Unfortunately, these readings did not provide any significant information but led to the other orders (up to the fourth order) of statistics being investigated [4]. From the first four orders of statistics, the mean proved to be the most informative and applicable for building a classifier [4]. In this paper, features obtained through applying the following procedures have been considered and are discussed in their respective subsections. They are the constant Q transform (CQT), power spectrum density (PSD) estimates, spectral centroid, spectral flatness measure (SFM), spectral flux, and features obtained through cepstral analysis.

a) *Constant Q Transform*

The CQT, as introduced by Brown in [10], yields a log-scaled time-frequency representation of the signal. It differs from the DFT in that the ratio between centre frequency and resolution remains constant making it suitable for the representation of musical signals as it improves time resolution as frequency increases.

b) *Spectral Centroid*

The spectral centroid is the 'centre of gravity' and is defined by the ratio of the sums of the magnitudes multiplied by the relevant frequencies all divided by the sum of magnitudes. It represents the 'brightness' of a signal and is calculated from the equation below [11]:

$$SC = \frac{\sum_{n=1}^{N-1} |X[n]| * f(n)}{\sum_{n=1}^{N-1} |X[n]|}$$

where N = length of the DFT

$|X(n)|$ = magnitude of the DFT

f = frequency at n

c) *Power Spectral Density*

The PSD describes the power distribution of the signal with respect to frequency [12]. Many methods exist for obtaining a PSD estimate and depending on the application, some are better suited than others. The periodogram is the simplest nonparametric method from which the PSD can be calculated. It is obtained directly from the signal itself by taking the FT of the autocorrelation of the windowed signal. However, it is not the most accurate method due to bias effects. This can be improved by selecting an appropriate windowing function. In this situation Welch's method, which is a nonparametric method, uses a Hamming window and provides a sufficiently detailed PSD. The straightforward periodogram uses a rectangular window.

d) *Spectral Flatness Measure*

The SFM is calculated from the power distribution via Welch's method and is defined as the PSD's geometric mean divided by its arithmetic mean [13].

$$SFM = \frac{geomean(PSD(windowed\_signal))}{mean(PSD(windowed\_signal))}$$

e) *Spectral Flux*

Spectral flux is a measure which represents the change in power between adjacent windows. It is obtained through the autocovariance of Welch's PSD of a windowed signal.

f) *Cepstral Analysis*

Cepstral analysis is a non-linear signalling technique often used in speech processing [12]. The real and Mel cepstra are considered in this paper. The real cepstrum is the inverse spectrum of the log of the spectrum. Whereas in the Mel cepstrum, which is a perceptually based spectrum, the data is converted into the Mel scale before the discrete cosine transform is carried out. Stages involved in obtaining the cepstra are shown in figure 1 below.

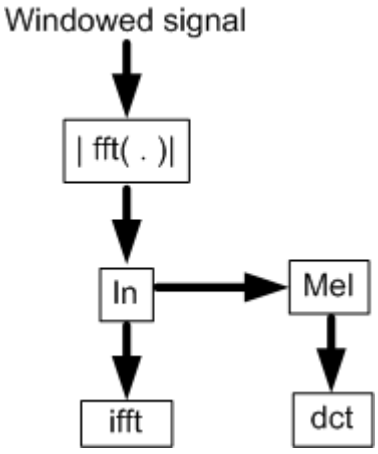

Figure 1: Steps involved in obtaining the real and Mel cepstra.

From these cepstra, the coefficients are obtained and the log energy of the signals is evaluated. The log energy is calculated from taking summing the natural logarithm of the magnitude of the FT of the signal and then by dividing this by the signal's length [14].

$$LogEnergy = \frac{sum(\log(abs(\,fft(signalwin))))}{length(signalwindow)}$$

VI RESULTS

a) *Constant Q Transform*

As can be seen in figure 2, due to the frequency resolution, the CQT domain is effective for visualizing and exploiting information about the harmonic content of a note.

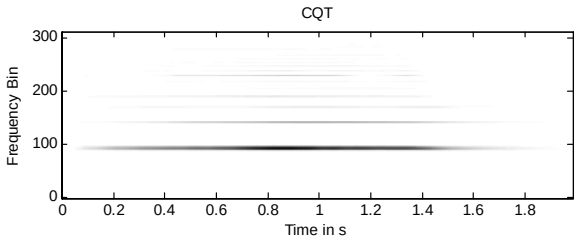

Figure 2: Harmonics visible via the CQT.

Based on the proportional strength of the strongest harmonic relative to the overall strength of all the harmonics in the signal, figure 3 clearly shows a significant difference between the beginner notes and the good player legato ones. This supports what professional stringed instrument players would say about beginners.

The proportional strength of harmonics has been calculated from the CQT by summing each frequency bin, taking the maximum and then dividing by the total.

$$\propto\_harm\_strength = \frac{\max\_freqbin\_value}{\sum all\_freqbins}$$

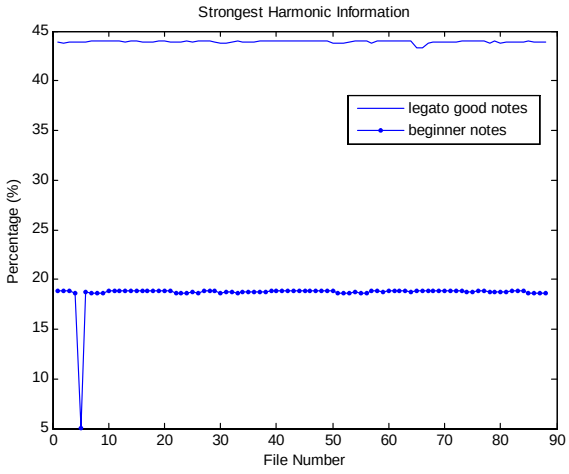

Figure 3: the proportional strength of harmonics obtained from CQT information.

b) *Spectral Centroid*

As a measure, it is more useful as a windowed measure from which the waveform can be split into regions (attack-steady-sate-decay). Tis can be seen in figure 4. However the spectral flatness measure (§VI.d) does this with much greater accuracy. The spectral centroid is better applied to

instrument identification tasks rather than within a timbre. As calculated, it is not sensitive enough a measure to be of use as a feature within the violin timbre space.

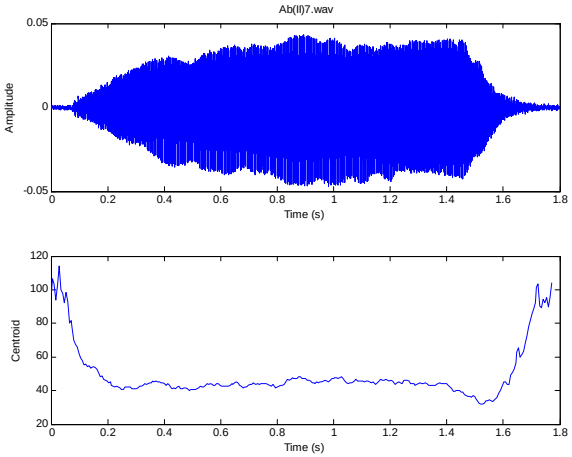

Figure 4: Waveform (top) and its moving spectral centroid (bottom).

c) *PSD*

The PSD from Welch's method is shown in the figure 5 below. A 1024 point Hamming window has been used with 50% overlap. Most of the energy is found at the fundamental frequency.

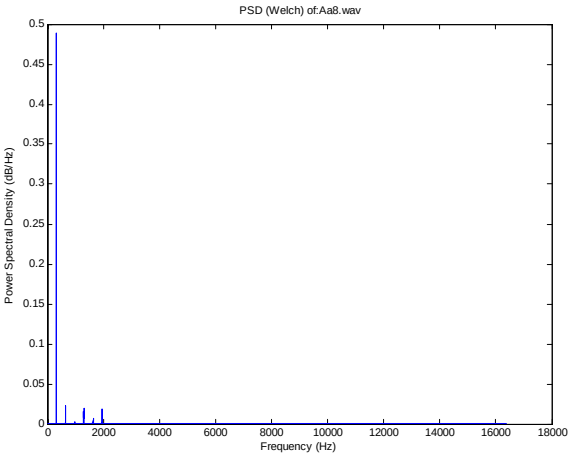

Figure 5: PSD via Welch's method.

d) *Spectral Flatness Measure*

Readings obtained from the SFM indicate how noisy or how close to a pure sinusoid a signal is. As the level approaches 1, the signal is closer to white noise. The closer to zero the reading, the closer the signal is to a pure sinusoid. This has proven to be very useful for sectioning real violin sounds. Figure 6 below compares a good legato note (top) with a beginner note (bottom).

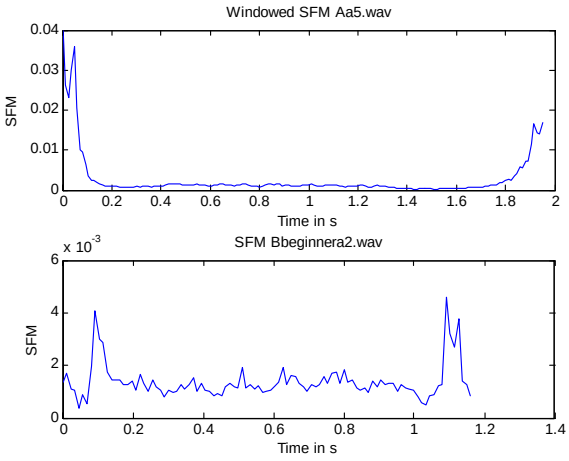

Figure 6: A moving SFM for a good legato note (top) and for a reasonable sounding beginner note (bottom).

The attack-steady-state-decay regions within the file become clear in the good note and are more approximate for the beginner note. These images hold much information about the bowing. The steepest changes occur at the beginning and ends of the note and this pattern is repeated throughout the good legato note files and reasonable sounding beginner files start approaching this shape too. The starts and ends of notes require more bow control than the middle section. These are also the regions where beginners typically 'crunch' due to lack of bow control. The pressure applied to the string via the bow is not kept the same throughout. The most pressure changes occur when the player in closest to either the tip (top of bow) or towards the heel (bottom of bow) and this is reflected in the SFM readings. The steady-state section of a good legato note, where pressure is applied more consistently, the SFM readings flatten out and approach zero. Attack, steady-state and decay sections become clear in figure 6, whereas obtaining this information from time or pitch methods is much more unreliable. This is important in that features can now be applied or developed according to region. For example, more accurate pitch detection can be carried out based only on the steady-state section of the waveform. This is important for string sounds as a significant acceptable fluctuation in pitch does exist due to the attack style and consequently physics of the string and instrument.

e) *Spectral Flux*

Disappointingly and not expected, spectral flux did not reveal useful results.

f) *Cepstral Analysis*

i. Cepstral Coefficients

Four orders of statistics were applied to the real and Mel cepstral coefficients. Mean, variance and kurtosis of the real cepstrum coefficients provided useful results for classification purposes as can be seen in figures 7, 8, and 9 respectively. Only the variance and kurtosis readings of the Mel cepstral coefficients, which are visible in figures 10 and 11 have shown to be useful. The mean did not separate the data lists in two distinct groups. The limitation of the real cepstrum is that it contains no phase information.

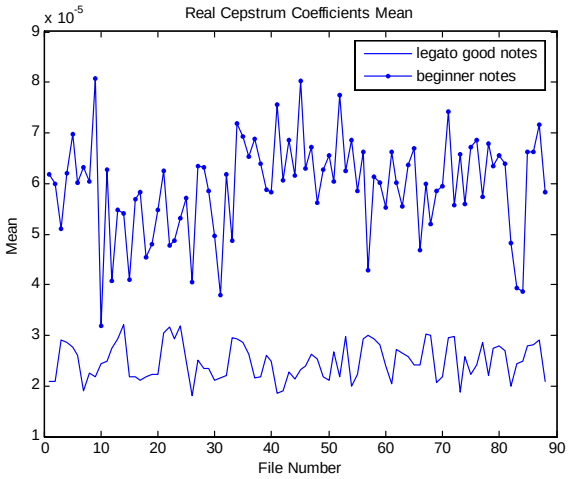

Figure 7: Mean values of the real cepstrum coefficients.

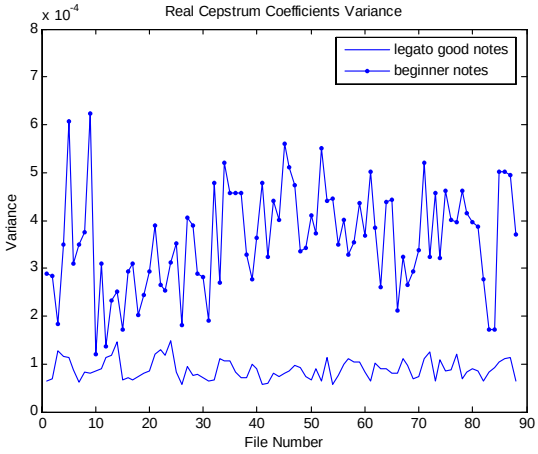

Figure 8: Variance readings of real cepstral coefficients.

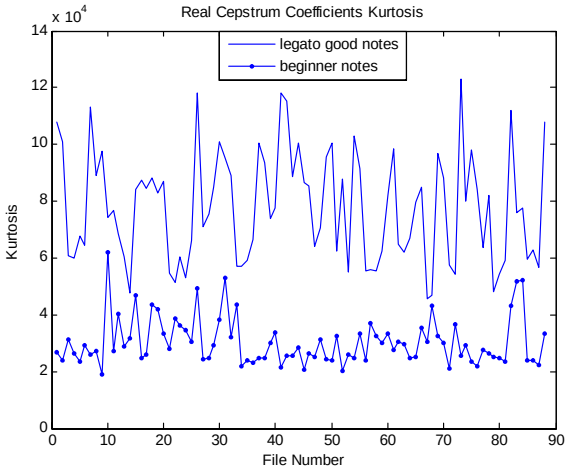

Figure 9: Kurtosis readings for real cepstral coefficients.

Converting into the Mel scale in this instance was not a distinct advantage. Developed by Stevens and Volkman, a Mel is a measure of perceived pitch of a tone [14]. It is not a linear scale and for this reason better represents the human auditory system. This could simply be due to the fact that all the data file pitches fall below 1 kHz. This is within the human speaking range which is the range where the human auditory system is at its most sensitive. However it is accepted that the real cepstrum provides the most successful results [14].

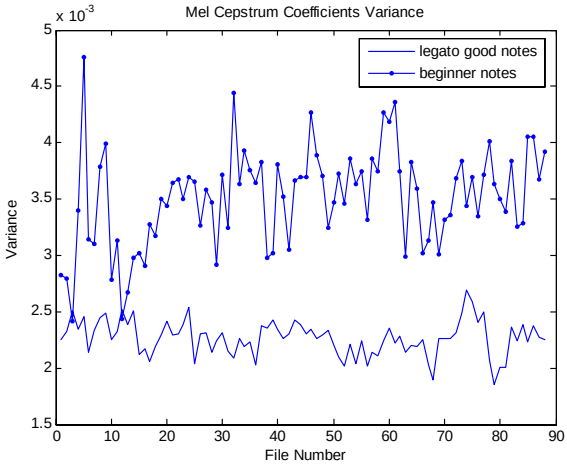

Figure 10: Variance readings for Mel cepstral coefficients.

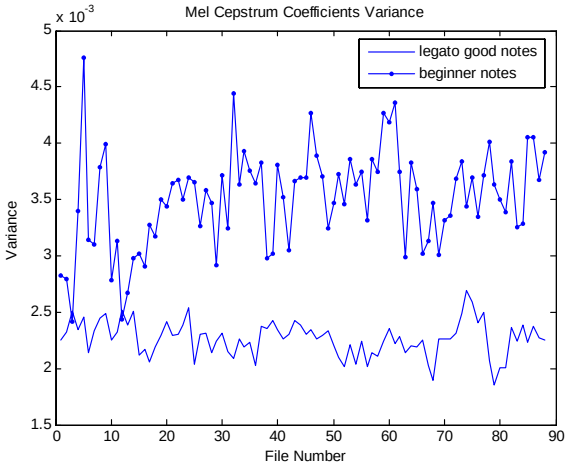

Figure 11: Kurtosis readings for Mel cepstral coefficients.

ii. Cepstral Log Energy

The log energy is often used as a relative measure of cepstral energy and how it changes [13]. Figures showing the log energy of the beginner notes versus good legato notes show distinct grouping patterns. It is also evident that the good legato notes have less variance and are more consistent which supports the fact that beginners have less bow control. As for beginners having higher energy readings, a logical explanation for this from a violinist's perspective is linked to efficiency and knowing how to make one's instrument resonate effortlessly.

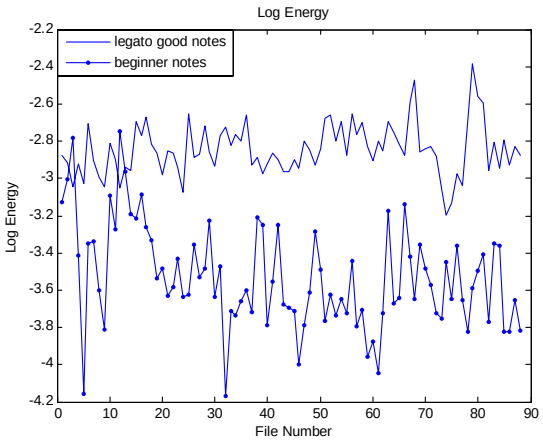

Figure: Real cepstral log energy.

VII CONCLUSIONS

The efficiency and usefulness of six features for describing timbre quality within the violin timbre space have been considered. Some of theses features work best on complete notes whereas others, such as the spectral flatness measure and the spectral centroid, are most effectively applied to a moving or windowed signal. The violin timbre space remains far from being defined in quantitative terms and work will be continued in this area.

VIII REFERENCES

[1] Charles, J. A., *et al.* 'Towards a Computer Assisted Violin Teaching Aid', International Symposium on Psychology and Music Education, Nov. 29-30, 2004, Padua, Italy.

[2] Eronen A., Klapuri, A. 'Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features', Signal Processing Lab, Tampere University of Technology, Tampere.

[3] Martin, K. D., Kim, Y. E., 'Musical instrument identification: A Pattern-Recognition Approach', 136th Meeting ASA, Oct. 1998.

[4] Charles, J. A. *et al.* 'Development of a Computer Based Teaching Aid, ViTool', AES 118th Convention, Barcelona, May 28-31, 2005.

[5] Hämäläinen, P., Mäki-Patola,T., Pulkki, V., Airas, M. 'Musical Computer Games Played by Singing', Proc. 7th Int. Conf. on Digital Audio Effects (DAFx'04), Naples, Oct. 5-8, 2004.

[6] Music Minus One, http://www.musicminusone.com/

[7] Meek, C., Birmingham, W. 'Johnny Can't Sing: A Comprehensive Error for Sung Music Queries', University of Michigan, Advanced Technologies Laboratory, 2002. ICMC

[8] Pollastri, E. 'Some Considerations About Processing Singing Voice for Music Retrieval', ISMIR 2002.

[9] Jackson, B. G., Berman, J., Sarch, K. *The A.S.T.A. Dictionary of Bowing Terms for String Instruments*, American String Teachers Association, 3rd edition, Tichenor Publishing Group, Bloomington, Indiana, 1987.

[10] Brown, J. C. 'Calculation of a Constant Q Spectral Transform', *Journal of the Acoustical Society of America*, 89, pp. 425-434, 1991.

[11] Beauchamp, J. W. 'Synthesis by Spectral Amplitude and 'Brightness' Matching Analyzed Musical Sounds', *Journal of Audio Engineering Society* 30(6), pp. 396-406, 1982.

[12] Oppenheim, A. V., Schafer, R. W. *Discrete-Time Signal Processing*, 2nd Ed., Prentice-Hall Int., 1999.

[13] Jayant, N. S., Noll, P. *Digital Coding of Waveforms*, Prentice Hall, Englewood Cliffs NJ, 1984.

[14] Deller, J. R., Hansen, J. H. L., Proakis, J. G. *Discrete-Time Processing of Speech Signals*, IEEE Press, John Wiley & Sons Inc., 2000.

[15] McAdams, S. 'Perspectives on the Contribution of Timbre to Musical Structure', *Computer Music Journal*, 23:3, pp. 85-102, 1999.