A Real-time Feedback Learning Tool to Visualize Sound Quality in Violin Performances

Sergio Giraldo, Rafael Ramirez, George Waddell, and Aaron Williamon

Universitat Pompeu Fabra (UPF), Barcelona, Spain
Centre for Performance Science, Royal College of Music, UK
{sergio.giraldo, rafael.ramirez}@upf.edu
{george.waddell, aaron.williamon}@rcm.ac.uk

Abstract. The assessment of the sound properties of a performed musical note has been widely studied in the past. Although a consensus exist on what is a good or a bad musical performance, there is not a formal definition of performance tone quality due to its subjectivity. In this study we present a computational approach for the automatic assessment of violin sound production. We investigate the correlations among extracted features from audio performances and the perceptual quality of violin sounds rated by listeners using machine learning techniques. The obtained models are used for implementing a real-time feedback learning system.

Keywords: Machine learning, Violin sound quality, Automatic assessment, Timbre dimensions, Audio features

1 Introduction

The quality of a performed sound is assumed to be a contribution of several parameters of sound such as pitch, loudness and timber. Eerola et al. (2012) identify 26 acoustic parameters of timbre among several instrument groups, that combined produce a particular sound quality, which might reflect a particular instrument performance technique, and/or the expressive intentions of the performer. Automatic characterization of dynamics and articulation from low level audio features has been studied by Maestre & Gómez (2005) in the context of expressive music performance. Knight et al. (2011) study the automatic assessment of tone quality in trumpet sounds using machine learning techniques. Romani Picas et al. (2015) make use of machine learning techniques to identify good and poor quality notes given training data consisting of low and high level audio features extracted from performed musical sounds. However, whereas pitch and dynamic measurements can be easily obtained from a computational perspective, a measure for timbre quality involves significant complications given that the exact formulation of timbre dimensions are still a matter of debate.

In this study we present an approach to automatically assess tone quality. Our aim is twofold: firstly, to understand the correlations between the proposed tone qualities, the ones previously used in the literature (e.g. Romani Picas

et al. (2015)) and the features extracted from the audio signal, and secondly, to generate machine learning models to predict the different proposed quality dimensions of the performance from the audio features. We have investigated the relationship between the terms that musicians use for quality assessment (e.g. clarity, warmth, depth, brilliance, resonance, richness, power) and low-level audio features (e.g. spectral centroid, spread, skewness, kurtosis, slope, decrease, roll-off point, flatness, spectral variation, spectral complexity, spectral crest MFCCs, and the energy in specific frequency bands), using machine learning techniques, based on the recordings and evaluations by music experts.

The predictive models were implemented and incorporated into a real-time feedback learning system, able to give automatic feedback about the timbral properties (Timbral dimensions) of exercises/notes being performed.

2 Methodology

2.1 Feature extraction and feature selection for real-time audio analysis.

Low and high-level audio features were extracted from the audio signals in both temporal and spectral domains using the Essentia library (Bogdanov et al. (2013)), using a frame size of 23ms, with a hop size of 11.5ms. On the other hand, perceptual tests to assess the quality of performed notes was conducted, in which 30 participants (with at least one year of musical training) were asked to mark sound quality in terms of predefined dimensions: dynamic, pitch and timbre stability, pitch accuracy and timbre richness, on a 7-point Likert scale. 27 Violin sounds were obtained from the public available data base by (Romani Picas et al. 2015), and selected in order to cover an homogeneous range of the violin's tessitura. Similarly, a proposed list of tone qualities (see Table 1), defined by music educators, was presented in pairs to the listener (e.g. Bright/Dark) to grade the sounds along a 7-point Likert scale.

Table 1: Proposed list of tone qualities by music experts

Tone Qualities	
Dark	Bright
Cold	Warm
Harsh	Sweet
Dry	Resonant
Light	Heavy
Grainy	Pure
Coarse	Smooth
Closed	Open
Restricted	Free
Narrow	Broad

2.2 Dynamics and intonation dimensions

Dynamics and pitch values were extracted from the audio by extracting the energy of the signal based on a frame-based calculation of the Root Mean Square (RMS), as well as, by obtaining frame based pitch values.

Pitch Accuracy (PA) . Pitch accuracy was measured in terms of the deviation in cents of the measured pitch to the closest tempered semitone. The actual pitch value was calculated in real-time on a frame basis (at 66 fps) using the Essentia Pitch Detection library. The obtained pitch values were smoothed using a 10 point average filter. Pitch accuracy was then obtained by obtaining the absolute difference between the pitch value and the closest semitone, and dividing by 50 cents (i.e. half of a semitone size). Thus, pitch accuracy ranges from 0 to 1, where the maximum deviation allowed is a semitone.

Pitch Stability (PS) . Pitch stability was calculated based on the standard deviation of the obtained frame-based value over a 300 ms historic window. Firstly, pitch frame-based values obtained with the Essentia Pitch Detection library were smoothed by applying a 10 point average filter. Standard deviation was calculated over a historical 300ms window. Low standard deviation values were assumed to indicate high pitch stability and vice versa.

Dynamic Stability (DS) . Dynamic stability was obtained by calculating the standard deviation of the energy over a 600 ms historic window. Firstly, we calculated a frame-based RMS (Root Mean Square) values. RMS values are later converted to decibel (dB) values and smoothed using a 10 point average filter. The standard deviation of the filtered RMS values was calculated over a 600 ms historical window. Similarly to pitch stability calculation, low standard deviation values were assumed to correspond to high dynamic stability and vice versa.

2.3 Timbral Dimensions Calculation

Timbral dimensions were calculated by training models which combined several of the audio features extracted with the Essentia library (Bogdanov et al. (2013)). Feature selection was performed over spectral descriptors, known to be close related to timbral characteristics of sound (see Peeters et al. (2011) for an overview), to obtain a subset of tonal descriptors that best predict each of the studied timbral dimension. The selected features include pitch, energy, spectral time-varying descriptors (centroid, spread, skewness, kurtosis, slope, decrease, rolloff, flatness, crest), spectro-harmonic (tristimulus 1, tristimulus 2, tristimulus 3, harmonic energy, noise energy). Mean and standard deviation over a 300 ms window was considered as well, for all the set of descriptors.

Timbre Stability (TS) and Timbre Richness (TR) The spectrum was obtained from the audio frame by means of the Fast Fourier Transform (FFT) and peak detection was performed on the spectrum afterwards. Based on the actual pitch value detected, the harmonic peaks were selected, allowing a 20% deviation from the ideal harmonic series. Later, spectral harmonic features, (e.g. tristimulus 1, 2 and 3) as well as time varying spectral features (e.g. kurtosis, skewness) were calculated.

2.4 Sound Dimensions Modelling.

Machine learning techniques were used to generate models to predict the different quality dimensions from the extracted features. Feature selection techniques were applied in order to obtain the subset of low level (frame-based) descriptors that best predict each of the studied sounds dimensions. Several machine learning schemes were compared, i.e., Linear Regression, M5-trees, Artificial Neural Networks, and Support Vector Machines.

Stability of pitch energy and timber Models were trained, to map the calculated standard deviations of the highest pitch stability rated sounds to 1 (good pitch stability) and, conversely, the bad examples to 0 (bad pitch stability). Correspondingly,models were trained to map the standard deviation values calculated in good/bad dynamic stability examples with a corresponding 0 to 1 dynamic stability value.

Timbre richness and stability Previously selected features were used to train models to order to best predict the ratings obtained on the surveys for timbral properties. For both Timbre Stability and Timbre Richness logistic regression models were obtained, using combinations of spectral features explained in Section 2.1.

3 Results

3.1 Tone survey

Consistency among participants ratings was assessed using Cronbach's coefficient (alpha). An acceptable degree of reliability was obtained (alpha>80, MCGraw and Wong, 1996) for all the sound examples. On the other hand, higher correlations (i.e. CC>0.8) were obtained between the overall quality of the sound and pitch stability/timbre richness.

3.2 Models accuracy

In Table 2 we present the Correlation Coefficient Index (CCI) obtained by the different models studied for the prediction of the rating on each of the dimensions considered. The obtained CCI of the models is presented as calculated in

both the Train Set (TS) and on a 10-Cross Fold validation scheme (CV), as an indicator of over-fitting. Consideration was also taken in terms of the feasibility of implementation of the models in a real-time application, giving priority to the ones less computationally expensive. For all the dimensions studied linear regression models were selected because of its overall good performance in terms of accuracy, low computational cost, and simplicity for implementation.

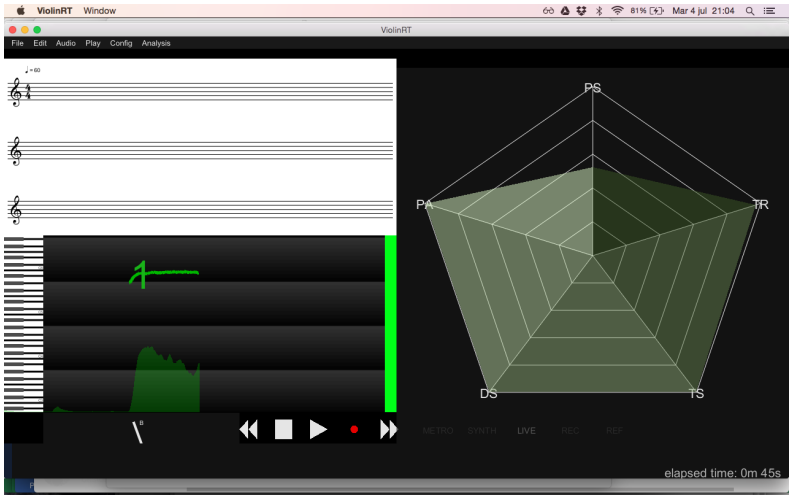
Table 2: Accuracies (CCI) for different sound's dimensions quality

Sound Dimension cv / train	Lin.Reg cv / train	Reg-Trees cv / train	SVMreg cv/ train	ANN
Pitch Accuracy	0.89 / 0.80	0.60 / 0.88	0.79 / 0.86	0.64 / 0.72
Pitch Stability	0.80 / 0.91	0.82 / 0.98	0.81 / 0.88	0.68 / 0.68
Dynamic Stability	0.82 / 0.84	0.67 / 0.87	0.78 / 0.85	0.69 / 0.65
Timbre Stability	0.80 / 0.89	0.63 / 0.91	0.86 / 0.81	0.60 / 0.75
timbre Richness	0.78 / 0.86	0.71 / 0.97	0.85 / 0.80	0.60 / 0.66

3.3 Real-time feedback learning widget

The aforementioned sound dimensions models for measuring the goodness in terms of the intonation, dynamics and tone, were presented in a intuitive graphic user interface, on the Violin RT app, as illustrated in Figure 1. Each sound dimension is presented on each axis of a spider chart, aiming at an intuitive user interaction in which the best sound quality is obtained when the chart is full filled.

Fig. 1: Real-time feedback system learning system screen shoot



4 Conclusions

In this paper a computational approach to automatically assess the quality of performed violin sounds was proposed. We conducted perceptual tests on the quality of recorded sounds based on previous defined quality dimensions, and studied the correlation among the different quality dimensions. Energy and spectral descriptors were extracted from the audio signal and machine learning models were obtained to predict the different quality dimensions from the audio features. Results indicate consistency among users responses, and the obtained models accuracy suggests that the extracted audio features contain sufficient information for characterizing the proposed tonal dimensions. Ongoing work includes extending the recording data, as well as modelling other tonal dimensions.

Acknowledgements

This work has been partly sponsored by the Spanish TIN project TIMUL (TIN2013-48152-C2-2-R), the European Union Horizon 2020 research and innovation programme under grant agreement No. 688269 (TELMI project), and the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., Serra, X. et al. (2013), Essentia: An audio analysis library for music information retrieval., *in* 'ISMIR', pp. 493–498.

Eerola, T., Ferrer, R. & Alluri, V. (2012), 'Timbre and affect dimensions: evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds', *Music Perception: An Interdisciplinary Journal* **30**(1), 49–70.

Knight, T., Upham, F. & Fujinaga, I. (2011), The potential for automatic assessment of trumpet tone quality., *in* 'ISMIR', pp. 573–578.

Maestre, E. & Gómez, E. (2005), Automatic characterization of dynamics and articulation of expressive monophonic recordings, *in* 'In Proceedings of the 118th Audio Engineering Society Convention', Citeseer.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N. & McAdams, S. (2011), 'The timbre toolbox: Extracting audio descriptors from musical signals', *The Journal of the Acoustical Society of America* **130**(5), 2902–2916.

Romani Picas, O., Parra Rodriguez, H., Dabiri, D., Tokuda, H., Hariya, W., Oishi, K. & Serra, X. (2015), A real-time system for measuring sound goodness in instrumental sounds, *in* 'Audio Engineering Society Convention 138', Audio Engineering Society.