A DATASET AND BASELINE FOR AUTOMATED ASSESSMENT OF TIMBRE QUALITY IN TRUMPET SOUND

Alberto Acquilino* **Ninad Puranik*** **Ichiro Fujinaga** **Gary Scavone**

Department of Music Research, Schulich School of Music, McGill University
Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)
555 Sherbrooke St W., Montreal, Canada H3A 1E3

alberto.acquilino@mail.mcgill.ca, ninad.puranik@mail.mcgill.ca

ABSTRACT

Music Performance Analysis is based on the evaluation of performance parameters such as pitch, dynamics, timbre, tempo and timing. While timbre is the least specific parameter among these and is often only implicitly understood, prominent brass pedagogues have reported that the presence of excessive muscle tension and inefficiency in playing by a musician is reflected in the timbre quality of the sound produced. In this work, we explore the application of machine learning to automatically assess timbre quality in trumpet playing, given both its educational value and connection to performance quality. An extensive dataset consisting of more than 19,000 tones played by 110 trumpet players of different expertise has been collected. A subset of 1,481 tones from this dataset was labeled by eight professional graders on a scale of 1 to 4 based on the perceived efficiency of sound production. Statistical analysis is performed to identify the correlation among the assigned ratings by the expert graders. A Random Forest classifier is trained using the mode of the ratings and its accuracy and variability is assessed with respect to the variability in human graders as a reference. An analysis of the important discriminatory features identifies stability of spectral peaks as a critical factor in trumpet timbre quality.

1. INTRODUCTION

The significance of tone quality in brass musical instruments has attracted considerable attention due to its relevance in areas such as pedagogy and musical performance. Teaching aural discrimination skills of tone quality is indeed a major component of music training [1]. The emphasis placed on the development of good tone quality can be attributed to its close relationship with sound production efficiency. In brass instrument pedagogy, there is a widely held belief that the most efficient sounds are perceived as rich and round, while less efficiently produced tones tend to sound strained and shrill [2–4]. This implies that a method that can accurately and consistently distinguish the perceived tone quality in a brass instrument may hold significant potential in pedagogical applications, providing guidance to beginning students on how to achieve greater efficiency in sound production. However, understanding factors that contribute to the timbral quality of trumpet sound remains an unsolved challenge thus far.

Playing a trumpet tone involves a complex interplay between the musician's embouchure, oral cavity, and airflow [5]. It is a delicate balance in which even the slightest alteration in any component contributing to the creation of a tone can result in changes to the overall timbre [6]. The multi-variable interaction that contributes to the characterization of timbre makes defining its quality a challenging task [7].

Helmoltz was among the first to attempt providing insight into the audio properties related to the quality of a musical tone by proposing a direct relationship to the quantity and to the relative intensity of its constituent partials [8]. In an exploratory study using the trumpet as a case study, Madsen and Geringer identified the amplitude of the first overtone as a discriminatory feature between tones of differing sound quality [9]. Building on this finding, a subsequent perceptual study by Geringer and Worthy analyzed the tonal quality of the trumpet by altering the content of partials in the sound [10].

In recent years, the investigation of trumpet tone quality has emerged as an area of inquiry within the field of Music Information Retrieval. A pioneering study conducted by Knight et al. examined the potential of a model classifier to categorize trumpet tones into two, three, and seven classes [11]. This research assessed 56 single- and multi-dimensional audio features, as well as their correlations with human judgments, utilizing a dataset comprised of 239 individual sounds. Despite the relatively low accuracy of the resultant model, this foundational work has paved the way for subsequent advancements in the automatic assessment of brass tone quality, highlighting its potential in pedagogical applications.

A subsequent collaborative research project between

* Equal contribution

the Music Technology Group of Pompeu Fabra University (MTG-UPF) and KORG Inc. employed machine learning algorithms to evaluate various musical parameters of trumpet sounds, including timbre quality [12, 13]. To the best of our knowledge, this represents the most recent investigation in this domain. The researchers collected and analyzed a publicly accessible dataset containing 738 trumpet sounds. However, the findings revealed a weak correlation between the scores generated by the trained model and the rankings assigned by human evaluators, indicating significant room for improvement in the model's performance. Limitations were also identified in relation to the reference dataset, which lacked diversity by utilizing sounds from only two graduated trumpet players, and in the proposed interface for implementation in pedagogical contexts [14].

The current study aims to provide a comprehensive exploration of this subject, incorporating a complete dataset of sampled sounds and expert-generated labels. [1] Section 2 describes dataset collection and preprocessing, while Section 3 presents the machine learning training, results and visualization based on the most important feature.

2. MATERIALS

The dataset employed for training the proposed model comprises auditory samples gathered by the first author at various music institutions and master classes throughout Europe before the start of his academic program at the host institution. In total, 110 distinct trumpet performers were recorded under varying acoustic conditions. To encompass the complete spectrum of sound production efficiency levels, individuals from diverse backgrounds were recorded, including students and instructors from amateur music schools, arts universities, orchestral musicians, and international jazz and classical soloists.

The same recording system was utilized across all data acquisition sessions, specifically the IM69D130 Shield2Go evaluation board developed by Infineon Technologies, which is equipped with two Infineon IM69D130 Micro-Electro-Mechanical Systems microphones. Such a microphone exhibits an Acoustic Overload Point of 130 dB, allowing it to capture loud audio signals such as those produced by a trumpet without distortion or saturation. Moreover, the microphone offers a sufficiently flat and extensive frequency response ranging from 20 Hz to 20 kHz, thereby covering the entire audible spectrum.

The selected evaluation board was connected to a Raspberry Pi 4 Model B and a Raspberry Pi Model 3B+ for recording. A sampling rate of 48 kHz and 32-bit depth were used for the acquisition of audio data. The subsequent section provides a detailed account of the recording methodology employed for audio data collection.

2.1 Dataset acquisition methodology

The data acquisition process involved inviting each musician into a room with a fairly low ambient noise level. A microphone was positioned approximately 50 cm in front of the trumpet bell and 10 cm from its longitudinal axis. In most instances, a set of two microphones was employed concurrently to ensure data redundancy, mitigating the risk of data loss should a device malfunction occur during the recording session.

Participants were instructed to play isolated tones of approximately one-second duration over a chromatic scale ranging from E3 to B♭5 at three distinct dynamic levels: *piano*, *mezzoforte*, and *forte*, in their preferred sequence. Musicians utilized their personal instruments and mouthpieces and were not required to adhere to a reference pitch (e.g., A4 at 440 Hz) as timbral quality concerning sound production efficiency is anticipated to be independent of a reference pitch.

The inclusion of various dynamic levels aimed to enhance the dataset's variability, as the timbre of brass instruments is significantly influenced by loudness [15]. A digital sound level meter was positioned adjacent to the microphone, providing real-time decibel level readings during the recording. Trumpet players were given indicative reference levels of 85 dB, 105 dB, and 115 dB, corresponding to the *piano*, *mezzoforte*, and *forte* dynamic levels, respectively.

Despite the specified guidelines, the dataset exhibits several inherent variabilities:

- The sustain duration of the tones ranged from 0.7 to 4 seconds.
- The chromatic scale's range was contingent upon the performer's skill level. Generally, less proficient musicians struggled to produce tones in the high register, in which case they were instructed to play up to their highest achievable note.
- For beginner musicians, playing a chromatic scale in front of a microphone proved challenging at times. Some participants opted to perform legato notes rather than separate tones.
- Less skilled musicians often experience difficulty in controlling the instrument's dynamic range, resulting in the recommended dynamic levels being primarily adhered to by more proficient players.

During the recording sessions, the first author, who holds a degree in trumpet performance and has professional experience as a musician and instructor, assigned a preliminary grade of the overall sound production efficiency on a scale of 1 to 100 to each player. Figure 1 illustrates the distribution of assigned grades divided into four ranges (i.e., 0–25, 26–50, 51–75, and 75–100), demonstrating that a substantial number of players are represented in each category.

The dataset under examination was partitioned into discrete trumpet tones utilizing the *pyin vamp* plugin developed by Mauch and Dixon [16], yielding a collection of over 19,000 tones. Although the segmentation process demonstrated a degree of inaccuracy, with certain audio

[1] The dataset can be accessed at: `https://github.com/PNinad/ISMIR2023`

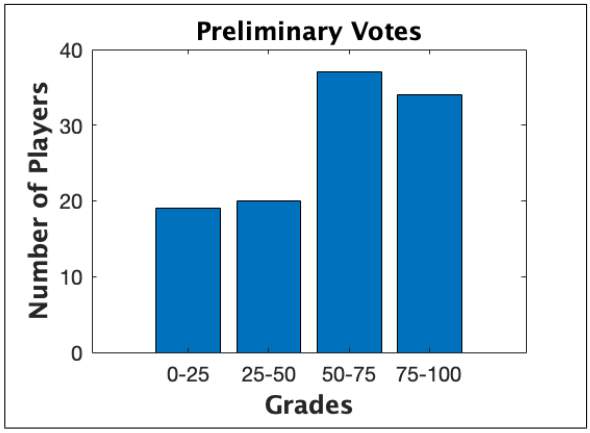

Figure 1. Distribution of recorded players according to the level of tone quality noted at the time of recording.

segments containing noise rather than trumpet tones, it nevertheless provided a satisfactory initial categorization of the data.

The following section outlines the methodology employed to prepare the dataset for label assignment by chosen evaluators.

2.2 Dataset preparation

Considering the approximate accuracy of the segmentation algorithm and the extensive nature of the overall dataset, it was decided to select a representative subset of the dataset for the manual examination of audio samples. To ensure that the whole range of tone quality is sufficiently represented, the subset was constructed of seventeen trumpet players such that five individuals had received a preliminary vote between 0–25 and four individuals with a grade between the other 3 ranges 26–50, 51–75, and 76–100 respectively. The first category was assigned one player more as the less experienced participants only partially cover the required chromatic scale, thus compensating for the lower representation of tones within this class. The selected subset encompassed 1,712 distinct trumpet tones.

It was decided to classify each tone into four categories based on their sound production efficiency, resulting in four classification levels: 1:poor, 2:fair, 3:good, and 4:excellent. This classification into four levels was employed with the intention of simplifying the label assignment process while retaining sufficient variability, as suggested by Wesolowski [17] and employed by Köktürk-Güzel et al. in a related research study [18].

The web interface shown in Figure 2 was subsequently developed to facilitate blind listening (i.e., without revealing the player's identity) and label assignment for each tone. The first author listened to all 1,712 sounds in the subset under analysis through the interface and assigned a label to each tone. The "Not a note" button enabled tagging of erroneously segmented sound samples which were filtered out to yield a dataset 1,481 clean samples.

The assignment of sound production efficiency class through anonymous listening to the audio samples in random order facilitated the allocation of a grade on a note-by-note basis, as opposed to providing an overall grade to the performance. This allowed for different grades to be assigned depending on the note if the level of sound efficiency varied along the chromatic scale. Additionally, the reliability of unbiased judgment could be assessed through a comparison with the preliminary grades assigned during the recording. The Spearman correlation coefficient between the two sets of grades was found to be 0.873 (P value<0.001), indicating the consistency of the author in assigning grades over time. This further indicates that players in general exhibit a consistent level of sound production efficiency along the chromatic scale.

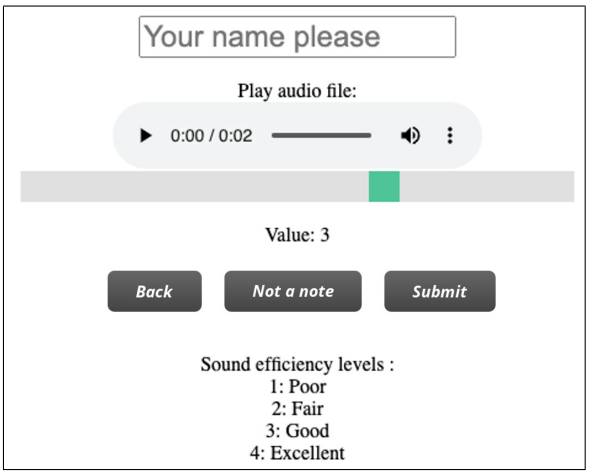

Figure 2. Interface for blind grading the trumpet tones.

2.3 Assessment labels

The cleaned dataset with 1,481 samples was subsequently presented to a panel of expert raters for evaluation via the described interface. A total of seven experts from different schools across Europe, North America and South America were chosen for the task. Among the raters, six were trumpet players, and one was a bass trombone player. All raters have professional experience as performers and/or teachers. This exploratory perceptual study was conducted online, with raters instructed to complete the task in a low-noise environment using professional headphones.

The rating sessions started with an introduction to the concept expressed by renowned brass instrument pedagogues, which asserts that rigidities in a trumpet player's body result in inefficiencies in playing, manifesting as a forced and strained sound. In contrast, a high-quality sound indicates efficiency of the embouchure and breathing muscles. Audio samples demonstrating extreme cases of this idea were presented and each rater confirmed their understanding of the concept and their ability to discern sound production efficiency in trumpet sounds based solely on audio information.

The dataset of 1,481 samples was split into two parts with 100 and 1,381 tones respectively. The raters first graded each of the 100 samples in approximately 15 minutes. After a 5 minute break, additional samples, randomly selected from the remaining 1,381 samples were presented

for evaluation. The raters continued to assess the trumpet tones until they experienced fatigue or until 90 minutes had elapsed from the beginning of the experiment. Table 1 displays the number of audio samples rated by each grader. Grader 1 corresponds to the first author who assigned the ratings manually by listening to all 1,481 samples in the subdataset, as described in the previous section. The set of 100 sounds were chosen such that they were equally distributed across the four classes, as determined from the labels by the author, and were used to ascertain the level of inter-rater reliability.

The next section describes the statistical analysis implemented on the data thus collected.

Grader ID	Graded tones
Grader 1	100 + 1381
Grader 2	100 + 401
Grader 3	100 + 206
Grader 4	100 + 312
Grader 5	100 + 383
Grader 6	100 + 366
Grader 7	100 + 564
Grader 8	100 + 491

Table 1. Number of individual tones evaluated by each grader.

2.4 Data analysis

The inter-rater reliability was assessed using the subdataset containing 100 tones graded by all the experts. Table 2 presents the Spearman ρ correlation coefficients with the corresponding P values for each pair of evaluators. As depicted in the table, all P values, representing the likelihood of obtaining the same results by chance, are less than 0.05.

The reported Spearman correlation coefficients range from 0.237 to 0.701. Notably, pairs including Grader 8 (the sole non-trumpet-playing expert) exhibited significantly lower correlation coefficients than all other pairs, potentially suggesting the significance of employing experts whose primary instrument aligns with the instrument under analysis for tasks of this nature. Due to the substantial differences in the ratings relative to the other raters, Grader 8 was deemed an outlier, and their results were excluded from further consideration. This adjustment increased Spearman ρ coefficients from 0.496 to 0.701, indicating fairly strong agreement among the judges [19].

Subsequently, a confusion matrix was computed for each evaluator, comparing the ratings assigned by that specific grader to the most frequently occurring (i.e., statistical mode) value in the ratings assigned by the seven evaluators for that specific tone. Cases where the mode was uncertain on one value were eliminated, resulting in 87 overall tones. The first seven subplots of Figure 3 display the resulting confusion matrices for each grader and their respective accuracy values (average f1 scores).

The next section describes the description of a model trained on the data obtained with reference to the variability of human assessment.

3. METHODOLOGY AND RESULTS

3.1 Audio Preprocessing and Model Training

The dataset preparation process described in Section 2.2 yielded a clean dataset with the audio samples of 1,481 tones. As a preprocessing step, the sound samples were first normalized to have a maximum signal amplitude equal to one. White noise at -60 dB was then added to the normalized audio to overcome the numerical errors (division by zero) encountered during feature extraction, without significantly altering the original signal. The audio features for each tone were then extracted using the Extractor algorithm from the Essentia library [20]. To reduce the computational complexity, only the statistical aggregates of the audio features (e.g., mean, variance, and mean of derivative) were utilized. Rhythm-based features were excluded since they were not deemed suitable for a timbre classification task. A total of 1,230 features were thus extracted to represent each audio sample.

As a first step, a Random Forest (RF) Classifier [21] was trained using the extracted audio features and labels provided by Grader 1, since Grader 1 had annotated each of 1,481 samples in the dataset. When the model was trained using the full set of audio features, a mean accuracy score of 78% was obtained in the 10-fold cross-validation. Using the model based feature selection in scikit-learn, the top 256 features were identified from an RF-classifier model trained using a 75%-25% train-test split of the dataset. Using just the top 256 features for training, the mean accuracy for the 10-fold cross-validation improved to 81.37%. The model thus obtained was implemented in a pedagogical application in a concurrent publication by the authors [22].

To eliminate the bias introduced by using a single grader, it was assumed that the most frequent label given by the expert graders is the true label. Only samples with at least two votes were used and samples which had equal number of votes for two labels by the expert graders were assumed to be ambiguous and were discarded from the dataset. With this approach, out of the 1,381 samples, 871 samples were deemed unambiguous. Similarly, 87 out of the 100 samples were unambiguous. An accuracy score of 59% was obtained on the test set of 87 samples for the RF model trained using the 871 samples as training set. The confusion matrices on the test-set for the different graders and the RF classifier can be seen in the bottom right subplot of Figure 3. It can be observed that most of the confusion is between the adjacent classes. Since the audio samples in the adjacent classes are in fact more similar to each other than the other classes, the errors seem to be reasonable, for both the graders and the model. While an accuracy score of 59% appears low, it is within the range of accuracy scores (53%–72%) of the human expert graders and it demonstrates that the extracted audio features could be used to classify the audio samples based on timbre quality.

The trained model was tested in real time by trumpet players and on labeled datasets different from the one in this study [12] showing promising generalisability.

Grader Pair	Grader						
	2	3	4	5	6	7	8
1	0.691*	0.668*	0.654*	0.645*	0.523*	0.638*	0.247***
2	-	0.701*	0.628*	0.650*	0.589*	0.650*	0.279**
3		-	0.599*	0.594*	0.496*	0.667*	0.237***
4			-	0.696*	0.650*	0.567*	0.349*
5				-	0.502*	0.637*	0.275**
6					-	0.524*	0.264**
7						-	0.353*

Table 2. Spearman ρ correlation coefficients between each pair of graders. Legend: * *p<.001*, ** *p<.01*, *** *p<.05*

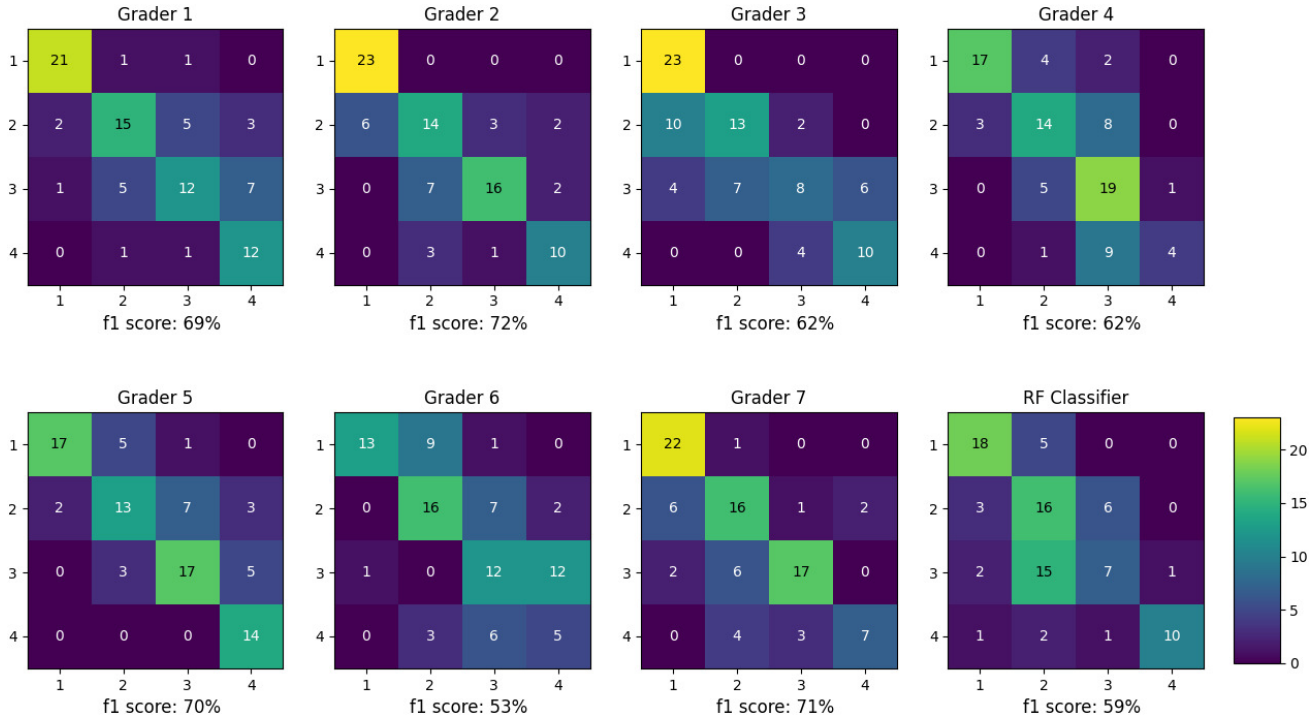

Figure 3. Confusion matrices with the predicted labels of each grader and of the trained RF classifier (horizontal axis) with respect to the true label as the mode of the assigned grade (vertical axis) and the corresponding f1 scores.

3.2 Feature importance

Due to a slightly subjective nature of the problem, there is considerable variability in the labels by human experts. Hence, very high classification accuracy scores cannot be achieved even with sophisticated machine learning models. However, even with a moderately accurate classifier, analysis of the most important features could help to develop an intuitive understanding of good quality timbre in trumpet sounds.

One of the main reasons to choose the Random Forest Classifier algorithm was that it gives access to the importance of each feature in the classification task. The feature importance scores for the classification are available as a model property in the scikit-learn implementation of the Random Forest algorithm [23]. The top 20 observed features are listed in Table 3.

Many of the top features are based on the mean of the derivative 'dmean' and the mean of the double derivative 'dmean2', suggesting that the change in the spectrum accross time is a crucial factor in the perception of the timbre quality. Notably three of the top features namely lowLevel.spectral_complexity.dmean, lowLevel.spectral_complexity.dmean2 and lowLevel.spectral_complexity.dvar are related to the time varying properties of the same underlying feature of spectral complexity.

A scatter plot of the lowLevel.spectral_complexity.dmean and lowLevel.spectral_complexity.dmean2 features considering only the best and worst class samples is shown in Figure 4. It is apparent that just this pair of features is quite successful in discriminating between the best and worst samples. Since both features are statistical aggregates of the spectral complexity feature, the raw feature was explored to develop a visualization of the sound production efficiency as described in the following subsection.

3.3 Visualization based on Spectral Complexity

Spectral complexity is based on the number of peaks in the spectrum of a time window [24]. The Essentia implementation of this feature considers the spectral peaks only up

Audio feature	Score (%)
lowLevel.spectral_complexity.dmean	1.381
lowLevel.scvalleys.mean_5	1.182
lowLevel.spectral_complexity.dmean2	1.049
lowLevel.spectral_complexity.dvar	0.897
lowLevel.sccoeffs.var_5	0.648
lowLevel.scvalleys.mean_3	0.636
lowLevel.sccoeffs.stdev_5	0.622
lowLevel.scvalleys.median_5	0.594
lowLevel.spectral_spread.dmean	0.570
sfx.tristimulus.dmean2_2	0.561
lowLevel.sccoeffs.median_4	0.531
lowLevel.sccoeffs.dmean2_3	0.496
lowLevel.scvalleys.median_3	0.492
lowLevel.barkbands.dmean_25	0.478
lowLevel.pitch_ instantaneous_confidence.dmean2	0.465
lowLevel.spectral_flux.dmean	0.465
lowLevel.spectral_complexity.dvar2	0.425
lowLevel.sccoeffs.mean_4	0.424
lowLevel.scvalleys.mean_2	0.412
lowLevel.spectral_complexity.stdev	0.402

Table 3. Top 20 features ranked by importance in the Random Forest Classifier.

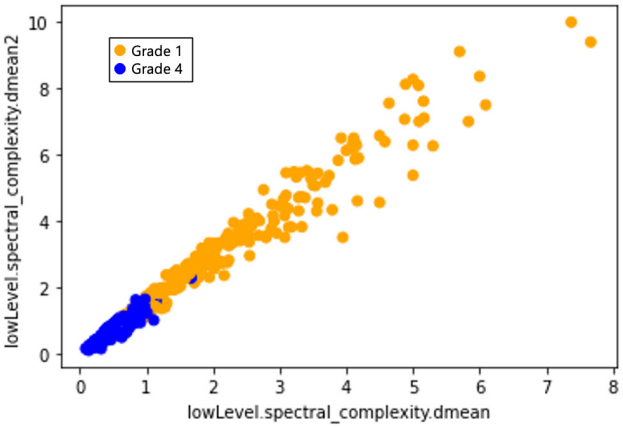

Figure 4. Scatter plot depicting the spectral complexity based features for best (blue) and worst (orange) class samples.

to 5 kHz. From the spectra of the collected dataset, the presence of harmonic peaks at frequencies higher than 5 kHz was evident. It was therefore decided to implement the spectral complexity considering the entire audible frequency range. To enhance peak detection accuracy, prior knowledge of the fundamental frequency 'f0' of the tone was utilized to search for spectral peaks exclusively in the vicinity of the integer multiples of the f0 frequency. For a normalized audio, peaks with signal energy less than -40dB were discarded to reduce noise. An FFT-bin mask was generated by assigning the value of one to the FFT bin if a peak was detected in it while all other bins were assigned a value of zero, thus generating a visualization to track the peaks across the analysis time windows.

Figure 5 shows the visualization for two representative sounds. It is evident that for sounds rated as excellent quality, the spectral peaks consistently lie in the same FFT-bin across time, leading to flat horizontal lines in the visualization. Whereas for sounds rated as poor quality, the spectral peaks show unsteadiness, particularly at the higher harmonics, which leads to broken and wavy lines in the visualization. The total number of peaks could be more or less depending on the f0 frequency of the note and the loudness. However, it appears that the perception of timbre quality is correlated to the steadiness of the peaks rather than their total number. A real-time implementation of this visualization could offer invaluable feedback on the efficiency of sound production, greatly benefiting new trumpet students who are still developing their auditory skills.

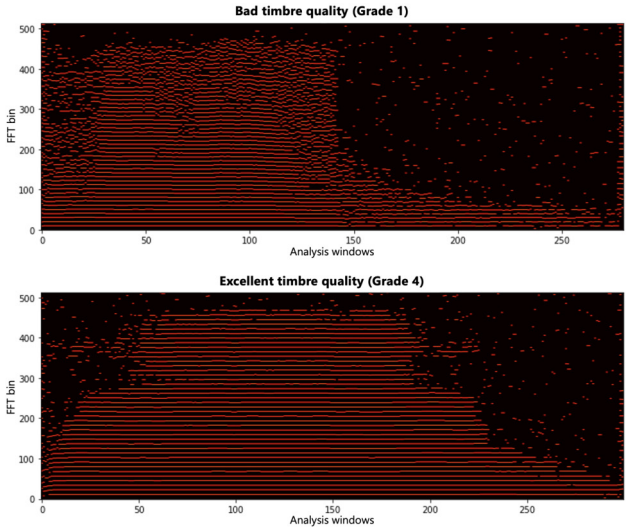

Figure 5. Visualization of the temporal evolution of spectral peaks for trumpet sounds rated as low-quality (top) and high-quality (bottom) timbre.

4. CONCLUSIONS

In this paper, we introduced the importance of timbre quality in trumpet performance and pedagogy. With an aim to develop an automated tool for the assessment and visualization of trumpet tone quality, an extensive dataset of trumpet tones was collected and manually graded with the help of experts. Through the inter-grader analysis presented, it was shown that while there are some differences in timbre preferences, most experts generally concur in differentiating the different levels of trumpet tone quality.

Random Forest Classifier models trained using extracted audio features were found to have accuracy scores comparable to the accuracy scores of human experts. Features based on spectral complexity were observed to have very high importance in the models trained for the task of trumpet timbre discrimination.

A representation based on the harmonic peaks in the spectrum was developed to visualize the timbre quality. The proposed visualization suggests that the stability over time of spectral partials plays an important role in discriminating the timbre quality of trumpet sounds.

Future research aims to incorporate the developed model and visualization in a pedagogical application and assess its efficacy in music classrooms.

5. ETHICS STATEMENT

Ethical approval for the study, including consenting procedures, was granted by the Research Ethics Board Office of McGill University following the guidelines of the Canadian Tri-Council Policy Statement.

Acknowledgments

This work was made possible with the support of a CIRMMT Student Award and a Tomlinson Doctoral Fellowship.

The authors thank the foundational contribution of Mirko d'Andrea and Emanuela Bussino for the audio data collection stages, as well as all the volunteers and colleagues who supported this research.

6. REFERENCES

[1] A. L. Simmons, "The relationship between prospective teachers' tone quality evaluations and their knowledge of wind instrument pedagogy," *Applications of Research in Music Education*, vol. 23, no. 2, pp. 42–51, 2005.

[2] A. Jacobs and B. Nelson, *Also Sprach Arnold Jacobs: A Developmental Guide for Brass Wind Musicians*. Polymnia Press, 2006.

[3] J. Thompson, *The Buzzing Book Complete Method; Trumpet or Other Brass Instruments*. Editions BIM, 2003.

[4] K. Steenstrup, *Teaching Brass*. Der Jyske Musikkonservatorium, 2007.

[5] F. G. Campos, *Trumpet Technique*. Oxford: Oxford University Press, 2005.

[6] S. Levarie and E. Levy, *Tone : A Study in Musical Acoustics. 2d ed.* Kent, Ohio: Kent State University Press, 1980.

[7] S. McAdams, S. Winsberg, S. Donnadieu, G. Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological Research*, vol. 58, pp. 177–912, Dec. 1995.

[8] H. von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York: Dover: Trans. By A. Ellis, 1977.

[9] C. Madsen and J. Geringer, "Preferences for trumpet tone quality versus intonation," *Bulletin for the Council for Research in Music*, vol. 46, pp. 13–22, 1976.

[10] J. M. Geringer and M. D. Worthy, "Effects of tone-quality changes on intonation and tone-quality ratings of high school and college instrumentalists," *Journal of Research in Music Education*, vol. 47, no. 2, pp. 135–149, 1999.

[11] T. Knight, T. Upham, and I. Fujinaga, "The potential for automatic assessment of trumpet tone quality," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2011, pp. 573–578.

[12] G. Bandiera, O. Romani, H. Tokuda, W. Hariya, K. Oishi, and X. Serra, "Good-sounds.org: A framework to explore goodness in instrumental sounds," in *Proceedings of the International Society for Music Information Retrieval Conference*, New York, 2016.

[13] O. Romani, H. Parra, D. Dabiri, H. Tokuda, W. Hariya, K. Oishi, and X. Serra, "A real-time system for measuring sound goodness in instrumental sounds," in *138th Audio Engineering Society Convention, AES*, Warsaw, Poland, 2015, pp. 1106–1111.

[14] A. Acquilino and G. Scavone, "Current state and future directions of technologies for music instrument pedagogy," *Frontiers in Psychology*, vol. 13, 2022.

[15] D. Luce and M. J. Clark, "Physical correlates of brass-instrument tones," *The Journal of the Acoustical Society of America*, vol. 42, pp. 1232–1243, 1967.

[16] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[17] B. C. Wesolowski, "Understanding and developing rubrics for music performance assessment," *Music Educators Journal*, pp. 36–42, 2012.

[18] B. E. Köktürk-Güzel, O. Büyük, B. Bozkurt, and O. Baysal, "Automatic assessment of student rhythmic pattern imitation performances," *Digital Signal Processing*, vol. 133, 2023.

[19] A. Williamon, J. Ginsborg, R. Perkins, and G. Waddell, *Performing Music Research: Methods in Music Education, Psychology, and Performance Science*. Oxford: Oxford University Press, 2021.

[20] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: An audio analysis library for music information retrieval," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2013, pp. 493–498.

[21] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[22] A. Acquilino, N. Puranik, I. Fujinaga, and G. Scavone, "Detecting efficiency in trumpet sound production: proposed methodology and pedagogical implications," in *Proceedings of the 5th Stockholm Music Acoustic Conference*, Stockholm, Sweden, 2023.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera, and X. Serra, "Indexing music by mood: design and integration of an automatic content-based annotator," *Multimedia Tools and Applications*, vol. 48, no. 1, pp. 161–184, 2009.