THE POTENTIAL FOR AUTOMATIC ASSESSMENT OF TRUMPET TONE QUALITY

Trevor Knight **Finn Upham** **Ichiro Fujinaga**
Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), McGill University
TrevorKnight@gmail.com Finn.Upham@gmail.com Ich@music.mcgill.ca

ABSTRACT

The goal of this study was to examine the possibility of training machine learning algorithms to differentiate between the performance of good notes and bad notes. Four trumpet players recorded a total of 239 notes from which audio features were extracted. The notes were subjectively graded by five brass players. The resulting dataset was used to train support vector machines with different groupings of ratings. Splitting the data set into two classes ("good" and "bad") at the median rating, the classifier showed an average success rate of 72% when training and testing using cross-validation. Splitting the data into three roughly-equal classes ("good," "medium," and "bad"), the classifier correctly identified the class an average of 54% of the time. Even using seven classes, the classifier identified the correct class 46% of the time, which is better than the result expected from chance or from the strategy of picking the most populous class (36%).

1. INTRODUCTION

1.1 Motivation

For some musical parameters, such as pitch or loudness, there are a well-established links between signal features of the audio file and perception [1]. Timbre is more complicated as several factors contribute to its perception [2]. The subjective quality of a musician's performance is more complicated still, with assumed contributions from pitch or intonation, loudness, timbre and likely other unknown factors [3].

The goal of this study is to determine the feasibility for computer analysis of performance quality. Given sufficient training data, is it possible for a computer to identify good and poor quality notes so as to give feedback to student musicians or for other pedagogical purposes.? This study also serves to create a dataset on which the signal components of tone quality may be examined.

The work was carried out by recording isolated notes played on trumpet by players with a range of experience, collecting subjective ratings of quality from human subjects, and training a classifier to identify note quality using extracted audio features. Because each of the notes were rated and analyzed in isolation, (i.e. as a single note without accompaniment or directed comparison), the note quality judgements in question are not likely to be affected by intonation, nor would they be related to other aspects of note quality dependent on musical context.

1.2 Tone Quality

Timbre is frequently defined as the differences between two sounds of the same pitch and loudness. This study was designed to isolate tone quality differences between notes of similar pitch, dynamics, and instrument. While numerous studies have attempted to determine the components of timbre that differentiate instruments and sounds [5-7], few studies have examined the auditory differences contributing to judgments of performance quality of tones. These studies most often use a technique called perceptual scaling to identify principal dimensions of timbre which generally aligned with the spectral content, the temporal change in the spectrum, and the quality of the attack [6,8]. With acoustically produced musical tones, however, these factors are interdependent and affect the perception of one another.

The contribution and inseparability of the different components of the sound is also found in pedagogical literature. In his instructional book on the trumpet, Delbert Dale says, "the actual sound of the attack (the moment the sound bursts out of the instrument) has a great deal to do with the sound of the remainder of the tone—at least to the listener" [9].

The few studies that have examined tone quality looked at specific aspects of the notes. Madsen and Geringer [4] examined preferences for "good" and "bad" tone quality in trumpet performance. Though the two tone qualities were audibly distinguishable when presented without accompaniment, the only difference their published analysis discussed was the amplitude of the second fundamental. In a different study, an equalizer was used to amplify or dampen the third through eleventh harmonics of recorded tones to be rated in tone quality [10]. For the brass instruments notes, a darker tone, caused by dampened harmonics, was judged to have a lower tone quality than the standard or brightened conditions.

The factors other than the amplitudes of the harmonics affect tone quality, and an examination of these is warranted. For the trumpet, tone quality is a product of the "balance and coordination" of embouchure, the oral cavity, and the airstream [11]. While "no two persons have the same or even similar tonal ideals" [9] and the standard for good and bad tone quality varies, common problems such as "a shrill piercing quality in the upper register, and a fuzzy and unclear tone in the lower register" [9] have been identified.

The goal of this study is to therefore see if it is possible to train a classifier that can use extracted audio features to make judgements about note quality consistent with average human judgements despite such variable and subjective criteria. The instructions given to our human participants (described later) are therefore intentionally vague to avoid biasing or limiting judgements and to avoid prescribing a definition of tone quality.

2. METHODS

2.1 Recordings

Recordings of the trumpet tones took place in a room designed for performance recording. The positions of the microphones, music stand, and player were the same for all recordings. Recordings were done using a cartioid microphone (DPA 4011-TL, Alleroed, Denmark) and a two channel recorder (Sound Devices 744T, Reedsburg, Wisconsin) at a bit depth of 24 and a sample rate of 48 kHz. The players had a range of experience and education on the trumpet. Player 1 is a musician whose primary instrument is the trombone and only played trumpet for this study. Player 2 is a trumpet player with twelve years of private lessons and regular ensemble performances at the university level both of which, however, ceased two years ago. Player 3 is currently an undergraduate music performance major who plays regularly with the university orchestra. Player 4 has been playing for 14 years with no instruction at the university level but with frequent live jazz performances.

The recorded phrases were three lines consisting of four half notes (minims) separated by half rests (minim rests). The same valve combination was repeated in the low range (A, Bb, B, C), mid range (E, F, F#, G), and high range (E, F, F#, G) and the players were instructed on which valves to use when a choice existed. Before recording each line, the players were given four clicks of a metronome at 60 bpm. The three lines were played with instructed dynamic levels of piano, then repeated at mezzo-forte and fortissimo.

With the exception of the trombone player, the musicians all recorded on their own trumpet and mouthpiece as well as a control trumpet (Conn Director, Conn-Selmer, Elkhart, Indiana) and mouthpiece (Bach 7C, Conn-Selmer). That is to say, three players recorded twelve notes at three different dynamic levels on two trumpets for a contribution of 214 notes. The trombone player, player 1, could not play the highest four notes and therefore contributed just eight notes at three dynamic levels on one trumpet for a total of 24 notes. One note from the dataset was excluded due to computer error so the total dataset had 239 notes.

2.2 Labeling

Individual notes were manually excised from the recordings to make discrete stimuli for subjective rating. Five brass players (three trumpet players, one trombone player, and one French horn player, all undergraduate or graduate music students with extensive performance experience) provided subjective labeling of the quality of the notes on a discrete scale from 1 to 7 with 1 labeled as "worst" and 7 labeled "best." The raters were instructed to listen to the note as many times as they wanted and to make a subjective rating of the note using anything they could hear and any criteria they deemed important, including their specific knowledge of brass instruments and the dynamic level. The notes were presented in three blocks (all the piano notes, all the mezzo-forte notes, all the fortissimo notes) but were randomized within each block.

Note quality judgements varied greatly per rater, as expected. While the intersubject ratings correlations averaged at r=0.50, some stimuli were rated more consistently than others. Dividing the 239 notes on the median standard deviation of 1.14 (on the discrete range of 1 to 7), the intersubject correlations on the more consistent subset of 118 (less than or equal to 1.14) averaged to r = 0.79. In contrast, the intersubject correlations on the remaining 121 stimuli averaged at r = 0.13, and failed to correlate significantly (i.e., with $p<0.05$) in 6 of 10 pair wise comparisons. Most of the bulge in the distribution of rounded average ratings, shown in figure 1, is due to these notes of ambiguous quality as they average to 4 or 5 with a couple dozen 3s and 6s. In the following analysis, all notes were represented only by their average rating across the five raters. The distribution of averaged ratings of the dataset is shown in Figure 1.

2.3 Feature Extraction

While studies have examined appropriate features for timbre recognition [12], timbre is just a subset of what potentially makes up the quality of a note. The extracted audio features were therefore widely selected, using 56 different features, of which 6 were multidimensional A complete list is given in the appendix. jAudio was used for feature extraction.[13]

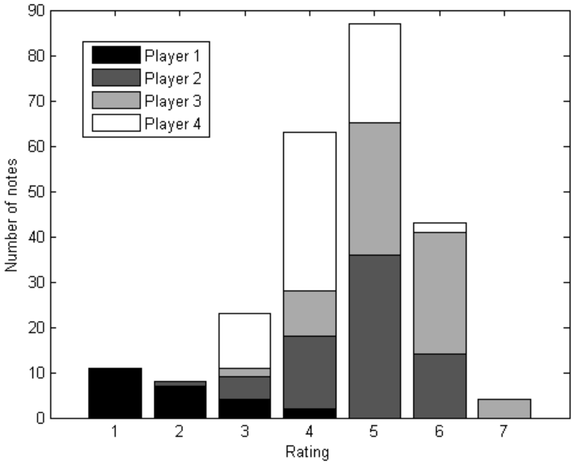

Figure 1. Histogram of the rounded average ratings from all raters and showing the contribution from each player.

2.4 Learning

2.4.1 Classifier Choice

ACE (Autonomous Classification Engine) 2.0, software used for testing, training, and running classifiers [14] was used throughout the study for these purposes. ACE was used to experiment with different classifiers including k-nearest neighbour, support vector machines (SVMs), several types of decision trees, and neural networks on a couple subsets of the data. .SVMs tended to perform best on these subsets.. For this reason and because of the relative interchangability of these techniques, SVMs were used throughout this study. In multi-class situations, however, SVMs do not encode an ordering of classes which makes the task slightly more difficult in the three and seven-class problems discussed below.

2.4.2 Groupings

Different groupings of the notes were used to test the accuracy of the classifiers, including two, three, and seven classes. While the judgments from the five raters were only integer values, each note was represented by a single average rating across all the raters and was therefore often a decimal number. The notes were assigned to classes based on this average rating.

Two-class problems were evaluated for three different groupings. The first grouping takes just the extremes of the data: the “good” class only has average ratings above 5.5 and the “bad” class has average ratings below 2.5, excluding all points in between. The second grouping is more inclusive, including all data below 3.5 for “bad” and above 4.5 for “good,” again excluding data in between. The last grouping includes all the data, split at the median rating, 4.6. The distribution of this labeling is shown in Figure 2.

Secondly, a grouping of three classes was also evaluated, splitting the data approximately into three groups, below 4.2, above or equal to 5.2, and the points in between.

Lastly, rounding the averaged ratings into the nearest category produced seven classes of data with labels 1 to 7. The distribution of this class is the same as seen in Figure 1.

2.4.3 Other tests

Furthermore, to test the performance of the classifier on notes from an unseen player we used a leave-one-player-out methodology. To do this, we repeated the above tests using three of the players to train and finding the success of classification on the fourth player. Because of the dominance of player 1 in ratings less than 2.5, we tested the seven class test with and without player 1 and did not test the two class problem using just the extremes of data (points less than 2.5 and greater than 5.5).

A classifier was also trained to test the possibility of discriminating between performers. To do this, each note was labeled only with a performer number, 1 through 4.

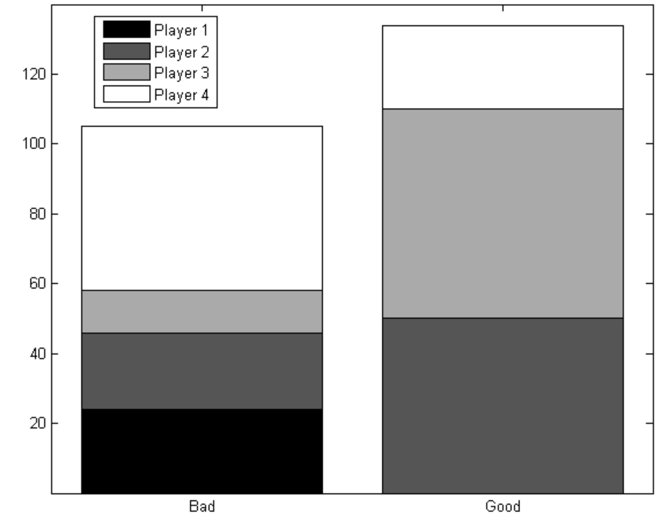

Figure 2: The distribution of the two classes when using all of the data, divided at the medan rating of 4.6.

3. RESULTS

For the two class problems, the most extreme data resulted in the highest success rate and increasing the inclusion of the classes lowered the average success of the five-fold cross validation. These results are summarized in Table 1.

For the three class problem, with a five-fold cross validation, an SVM correctly identified the class on average 54.0% of the tones. This result is shown in Table 2.

"Bad"		"Good"		Average Success
Range	Number	Range	Number	
1–2.4	19	5.6–7	47	96.9%
1–3.4	42	4.6–7	134	87.5%
1–4.5	105	4.6–7	134	72.0%

Table 1: Classifier results with two classes and five-fold cross validation

"Bad"		"Middle"		"Good"		Average Success
Range	Number	Range	Number	Range	Number	
1–4.1	77	4.2–5.1	86	5.2–7	76	54.0%

Table 2: Classifier results with three classes and five-fold cross validation

The five-fold cross-validation success of the seven class problem is shown in Table 3 and the confusion matrix is shown in Table 4. The rows labels represent the true classifications of the instances and the columns labels are the classifications assigned by the SVM. For instance, of the notes of class 1, eight were correctly identified but one note was labeled 3 and two were labeled 4.

Class	1	2	3	4	5	6	7	Avg. Success
Number	11	8	23	63	87	43	4	46.03%

Table 3: Classifier results with seven classes and five-fold cross validation

	1	2	3	4	5	6	7
1	8		1	2			
2	2			4	2		
3			1	15	6	1	
4				26	35	2	
5				22	56	9	
6				3	21	19	
7					1	3	

Table 4: The confusion matrix for the seven-class problem; the correct classes are given in the row labels.

When using the leave-one-player-out test, the success rate decreased. A summary is shown in Table 5.

For the performer identification task, with five folds, the classifier averaged 88.3% success. The confusion matrix is shown in Table 6. Again the correct label is the row label. For example, player one played 24 notes, of which 21 were identified correctly, two were incorrectly labeled as player 2 and one labeled as player 3.

Player tested					
1	2	3	4	Avg.	
23%	66%	84%	67%	**60%**	2 classes (1–3.5, 4.6–7)
67%	60%	47%	51%	**56%**	2 classes (split at 4.6)
58%	35%	39%	38%	**42%**	3 classes
0%	25%	24%	38%	**22%**	7 classes
	26%	25%	39%	**30%**	7 classes (w/o player 1)

Table 5: Results for leave-player-out classification.

	1	2	3	4
1	21	2	1	
2	1	61		10
3		1	68	3
4	3	5	2	61

Table 6: The player identification confusion matrix; the correct player identifications are given by the row-labels.

4. DISCUSSION

The classifiers show a surprising ability to discriminate between classes based on the extracted features with two, three, and seven classes. Even with seven classes, the classifier identified the correct class 46% of the time, which is better than chance or the success rate expected from picking the most common class (36%). This shows promise for the possibility to train a classifier to give automatic feedback on student musicians' performance.

There are, however, severe limitations to this data set. Because there are only four players in the data set, each with a distinct distribution of notes, there may be latent features unrelated to performance quality that can help narrow the selection of class and improve classifier success. This hypothesis is bolstered by the high success in performer identification task. For comparison, a 1-note attempt at identifying the correct performer out of three possible performers gave at best a 43% success in a previous study [15].

The classifier's success with the subset of 118 notes with rating standard deviation less than or equal to 1.14 was not different than the dataset as a whole. This seems to indicate the classifier is not using the same cues or salient features that allowed or encouraged agreement between the raters.

The results for the leave-one-player-out task decreased sharply compared to the result using all players and testing with cross-validation. This could be because of the distinct distribution of each player and/or other distinct features that identify one performer compared to another.

In the seven class identification task, mathematically, for a note to be considered of class one (or 7) there had to be strong agreement among the raters, as at least 3 of the raters had to rate that note as class one. This distinctively bad performance of class 1 notes probably led to the relatively high success in identifying them (8 out of 11 correct) compared to, for example, class 2 which had no correct identifications. As well, because player 1 was not able to record the top four notes of the exercise, having a higher pitch note skews the rating towards the upper end of ratings.

Further work is needed to examine the robustness of these results with more players and with different recording conditions, such as notes of varying duration, or using phrases of several notes.

5. ACKNOWLEDGEMENTS

This work was made possible by a CIRMMT Student Award and the amazing help of the Harold Kilianski, Yves Méthot and Julien Boissinot of CIRMMT. Partial funding for the work was also provided by Fonds Québécois de la Recherche sur la Société et la Culture (FQRSC) and Social Sciences and Humanities Research Council of Canada (SSHRC).

6. APPENDIX: FEATURES EXTRACTED

Beat Sum Overall Average
Beat Sum Overall Standard Deviation
Compactness Overall Average
Compactness Overall Standard Deviation
Derivative of Partial Based Spectral Centroid Overall Average
Derivative of Partial Based Spectral Centroid Overall Standard Deviation
Derivative of Root Mean Square Overall Average
Derivative of Root Mean Square Overall Standard Deviation
Derivative of Spectral Centroid Overall Average
Derivative of Spectral Centroid Overall Standard Deviation
Derivative of Spectral Flux Overall Average
Derivative of Spectral Flux Overall Standard Deviation
Derivative of Spectral Rolloff Point Overall Average
Derivative of Spectral Rolloff Point Overall Standard Deviation
Derivative of Strongest Frequency Via Zero Crossings Overall Average
Derivative of Strongest Frequency Via Zero Crossings Overall Standard Deviation
Fraction Of Low Energy Windows Overall Average
Fraction Of Low Energy Windows Overall Standard Deviation
LPC Overall Average
LPC Overall Standard Deviation
Method of Moments Overall Average
Method of Moments Overall Standard Deviation
MFCC Overall Average
MFCC Overall Standard Deviation
Partial Based Spectral Centroid Overall Average
Partial Based Spectral Centroid Overall Standard Deviation
Root Mean Square Overall Average
Root Mean Square Overall Standard Deviation
Spectral Centroid Overall Average
Spectral Centroid Overall Standard Deviation
Spectral Flux Overall Average
Spectral Flux Overall Standard Deviation
Spectral Rolloff Point Overall Average
Spectral Rolloff Point Overall Standard Deviation
Spectral Variability Overall Average
Spectral Variability Overall Standard Deviation
Standard Deviation of Compactness Overall Average
Standard Deviation of Compactness Overall Standard Deviation
Standard Deviation of Partial Based Spectral Centroid Overall Average
Standard Deviation of Partial Based Spectral Centroid Overall Standard Deviation
Standard Deviation of Root Mean Square Overall Average
Standard Deviation of Root Mean Square Overall Standard Deviation
Standard Deviation of Spectral Centroid Overall Average
Standard Deviation of Spectral Centroid Overall Standard Deviation
Standard Deviation of Spectral Flux Overall Average
Standard Deviation of Spectral Flux Overall Standard Deviation
Standard Deviation of Strongest Frequency Via Zero Crossings Overall Average
Standard Deviation of Strongest Frequency Via Zero Crossings Overall Standard Deviation
Standard Deviation of Zero Crossings Overall Average
Standard Deviation of Zero Crossings Overall Standard Deviation
Strength Of Strongest Beat Overall Average
Strength Of Strongest Beat Overall Standard Deviation
Strongest Frequency Via Zero Crossings Overall Average
Strongest Frequency Via Zero Crossings Overall Standard Deviation
Zero Crossings Overall Average
Zero Crossings Overall Standard Deviation

7. REFERENCES

[1] R. Plomp, *Aspects of Tone Sensation: A Psychophysical Study*, New York, NY: Academic Press, 1976.

[2] 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, Dec. 1995, p. 177–92.

[3] J. Geringer and C. Madsen, "Musicians' ratings of good versus bad vocal and string performances," *Journal of Research in Music Education*, vol. 46, 1998, p. 522–34.

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

[5] S. McAdams and J.-C. Cunible, "Perception of Timbral Analogies," *Philosophical Transactions: Biological Sciences*, vol. 336, 1992, p. 383–9.

[6] C. Krumhansl, "Why is musical timbre so hard to understand?," *Structure and Perception of Electroacoustic Sound and Music: Proceedings of the Marcus Wallenberg Symposium*, Lund, Sweden: 1988, p. 43–53.

[7] P. Iverson and C. Krumhanslm, "Isolating the dynamic attributes of musical timbre," *Journal of Acoustical Society of America*, vol. 94, 1993, p. 2595–603.

[8] S. Handel, "Timbre perception and auditory object identification," in *Hearing*, B. Moore, ed., San Diego: Academic Press, 1995, p. 425–61.

[9] D. Dale, *Trumpet Technique*, London: Oxford University Press, 1975.

[10] J. Geringer and M. 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, Jan. 1999, p. 135–49.

[11] F. Campos, *Trumpet Technique,* New York: Oxford University Press, 2005.

[12] X. Zhang and W.R. Zbigniew, "Analysis of sound features for music timbre recognition," *International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, IEEE, 2007, p. 3–8.

[13] C. McKay, I. Fujinaga, and P. Depalle, "jAudio: A feature extraction library," *Proceedings of the International Conference on Music Information Retrieval*, 2005, p. 600–3.

[14] J. Thompson, C. Mckay, J.A. Burgoyne, and I. Fujinaga, "Additions and improvements to the ACE 2.0 music classifier," in *Proceedings of the International Conference on Music Information Retrieval*, 2009.

[15] R. Ramirez, E. Maestre, A. Pertusa, E. Gomez, and X. Serra, "Performance-based interpreter identification in saxophone audio recordings," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, Mar. 2007, p. 356–64.