Analysis of Objective Descriptors for Music Performance Assessment

Siddharth Gururani, Kumar Ashis Pati, Chih-Wei Wu and Alexander Lerch

Center for Music Technology, Georgia Institute of Technology, USA

{sgururani3, apati3, cwu307, alexander.lerch}@gatech.edu

Abstract

The assessment of musical performances in, e.g., student competitions or auditions, is a largely subjective evaluation of a performer's technical skills and expressivity. Objective descriptors extracted from the audio signal have been proposed for automatic performance assessment in such a context. Such descriptors represent different aspects of pitch, dynamics and timing of a performance and have been shown to be reasonably successful in modeling human assessments of student performances through regression. This study aims to identify the influence of individual descriptors on models of human assessment in 4 categories: musicality, note accuracy, rhythmic accuracy, and tone quality. To evaluate the influence of the individual descriptors, the descriptors highly correlated with the human assessments are identified. Subsequently, various subsets are chosen using different selection criteria and the adjusted R-squared metric is computed to evaluate the degree to which these subsets explain the variance in the assessments. In addition, sequential forward selection is performed to identify the most meaningful descriptors. The goal of this study is to gain insights into which objective descriptors contribute most to the human assessments as well as to identify a subset of well-performing descriptors. The results indicate that a small subset of the designed descriptors can perform at a similar accuracy as the full set of descriptors. Sequential forward selection shows how around 33% of the descriptors do not add new information to the linear regression models, pointing towards redundancy in the descriptors.

I. Introduction

A musical performance by a human requires the interpretation of musical ideas, typically from a score, the planning of retrieved musical units and transforming thoughts into motion. This makes the task one of the most complex serial actions performed by a human (Palmer, 1997). Therefore, students learning to perform music require regular feedback and attention from trained teachers. This feedback may be in the form of qualitative and quantitative assessment of the student performances on various criteria such as rhythmic and pitch accuracy.

A common issue with such feedback, however, is its highly subjective nature (Wesolowski, 2012). This may include bias and lead to inconsistencies counter-productive to the students' learning experience. Thus, computational methods for music performance assessment are sought after because they are able to provide objective, consistent, and reproducible assessments.

Tools and techniques for automatically extracting musical information from audio signals matured as a result of advances in music information retrieval (MIR). Relevant tasks are pitch and beat tracking, transcription and source separation in audio signals. These MIR tasks may be treated as fundamental building blocks for automatic music performance assessment. Several academic works leverage these techniques to develop automatic music performance assessment systems (Dittmar, Cano, Abeßer, & Grollmisch, 2012). In addition, companies such as Smart Music (http://www.smartmusic.com Last access: 2018/05/26) and Yousician (https://get.yousician.com Last access: 2018/05/26) have developed commercial software for performance assessment.

Typical automatic performance assessment systems comprise of algorithms that extract descriptors or features from the audio signal. These descriptors are, in turn, used to train statistical models using expert assessment data. The models are then applied to predict the assessment scores. In this paper, we focus on the importance of these descriptors for modelling student performances.

The structure of the paper is as follows: Section 2 provides a brief overview of related work. Section 3 introduces the descriptors and the dataset used for the analysis. The methodology and experimental setup are described in Section 4. The results of the experiments and conclusion follow in Sections 5 and 6 respectively.

II. Related Work

Objective descriptors, also known as audio features, are quantities that represent various characteristics of an audio signal. These are typically computed on short blocks of the audio signal to capture short-term characteristics and then summarized using statistical measures such as the mean and standard deviation (Lerch, 2012). In the context of music performance assessment, different descriptors may be used to capture low-level information pertaining to signal energy and timbre and subsequently linked to high-level semantic concepts through computational models.

According to work done by Vidwans et al. (Vidwans et al., 2017), objective descriptors for music performance may be broadly categorized into two major categories: (i) Score-independent and (ii) Score-dependent descriptors. Score-independent descriptors are derived without using additional information about the musical score being performed. The benefit of using these descriptors is that they only require the audio file and do not rely on the availability of the score. The intuition behind using this approach is that humans are able to assess performances even without the score by observing pitch and rhythm stability, among others. Some examples of work involving score independent features are (Abeßer, Hasselhorn, Dittmar, Lehmann, & Grollmisch, 2013), (Han & Lee, 2014) and (Wu et al., 2016).

Score-dependent descriptors are derived by leveraging the information provided by the score being performed. The advantage of using this approach is that direct comparison between the performance and the score is possible and therefore, more accurate descriptors of the performance may be extracted. These are applicable in the audition setting where assessors have access to the score that is supposed to be performed. Work investigating score-dependent features are (Vidwans et al.,

2017), (Devaney, Mandel, & Fujinaga, 2012) and (Mayor, Bonada, & Loscos, 2009).

In addition to the aforementioned categories, objective descriptors can also be automatically inferred from the data using machine learning techniques such as sparse coding (Wu & Lerch, 2018) or neural networks (Pati, Gururani, & Lerch, 2018). Learned features allow the extraction of relevant information that might be overlooked by human engineers, however, these features tend to be abstract and have no specific physical meaning. Thus, they are outside the scope of this paper.

In previous work by (Wu et al., 2016) and (Vidwans et al., 2017), score-independent and score-dependent descriptors were designed and shown to be useful for the task of automatic music performance assessment. These studies trained regression models using the large set of descriptors to achieve the best performance. However, an analysis of the importance or contribution of the descriptors is not performed. To increase the interpretability of such approaches and gain more insights about the system, we aim to analyze these descriptors in detail using various methods and identify the well-performing set of descriptors from among the larger set of descriptors.

III. Dataset and Descriptors

Dataset

The dataset used in this paper is obtained from the Florida Bandmasters Association (FBA). It consists of audio recordings of All-State auditions of middle and high school students. Each recording consists of exercises such as etudes, scales, and sight reading and is accompanied by expert assessments in the four following categories: musicality, note accuracy, rhythmic accuracy and tone quality. For more details about the dataset, we refer readers to our previous work, (Wu et al., 2016) and (Vidwans et al., 2017). We consider middle school students performing alto saxophone (n = 392). Only the technical etude is considered for these experiments.

Descriptors

The descriptors investigated here have shown their meaningfulness in previous studies. They were designed to model different facets of a student performance. We provide a brief overview of the descriptors used in this paper and refer readers to work done by (Vidwans et al., 2017) for a detailed description of all the descriptors designed for this task.

As described in Section 2, the descriptors chosen are broadly categorized into two classes:

1. Score-independent
2. Score-dependent

The score-independent descriptors may be further divided into 3 categories:

Pitch: Descriptors extracted from the pitch contour of the performance fall under this category. They include measures for note steadiness, accuracy and intonation. They are computed on a note-by-note basis and aggregated for an entire performance using the mean, standard deviation, maximum and minimum value.

Rhythm: Descriptors extracted from the inter-onset-interval (IOI) histogram computer from note onset times. They measure the timing accuracy of the note. Standard statistical measures are extracted from the histogram such as crest, skewness, rolloff, etc.

Table 1. Score-Independent Descriptors

Index	Descriptor	Description
N1	Pitch 1	Average note accuracy
N2-5	Pitch 2	St. dev. of pitch values (mean, st. dev., min, max)
N6-9	Pitch 3	% of pitch values deviating more than one st. dev. (mean, st. dev., min, max)
N10	Intonation	% of notes in tune
N11-14	Dynamics 1	Amplitude deviation (mean, st. dev., min, max)
N15-18	Dynamics 2	Amplitude envelope spikes (mean, st. dev., min, max)
N19-24	Rhythm	Crest, bin resolution, skewness, kurtosis, roll-off, power ratio of the IOI histogram

Dynamics: Descriptors extracted from the amplitude of each note. These include note-level descriptors that measure amplitude steadiness or spikes. Similar to pitch descriptors, they are aggregated using the mean, standard deviation, maximum and minimum across all notes.

Score-dependent descriptors are computed after aligning the pitch sequence or contour of the performance with the score of the performance. Alignment gives a more accurate segmentation of the notes in the performance. The score-dependent descriptors are also of 3 types:

Pitch: Most of the pitch descriptors are similar to the score-independent descriptors and differ in the fact that note boundaries are computed using score alignment. In addition, we compute descriptors measuring the deviation of the played note from the note in the score.

Rhythm: Similar descriptors as the score-independent are computed with score-aligned note onsets.

Alignment: The score alignment is performed using dynamic time warping (Müller, 2007) of the pitch contour with the sequence of notes in the score. Descriptors are extracted from the alignment path such as the length and deviation of the slope from a straight line. In addition, the cost of alignment and a measure of extra notes or unplayed notes are computed.

There are 46 descriptors that are analyzed in this paper. 24 of these are score-independent and 22 are score-dependent. We index these descriptors prefixed by 'N' for score-independent and 'S' for score-dependent descriptors. Tables 1 and 2 enumerate all the descriptors and their index.

IV. Experiments

In this section, we describe the two experiments carried out to analyze the importance of the extracted descriptors. The first experiment involves studying the direct correlation between the descriptors and the assessments. The second involves different methods for selecting descriptor subsets and subsequently using these subsets to train linear regression models to predict the assessments.

Correlation Analysis

In this experiment, we aim to study how correlated or decorrelated each of the descriptors is with the human assessments. Since these descriptors have been used to model human assessments, we investigate whether a relation can be

Table 2. Score-dependent Descriptors

Index	Descriptor	Description
S1	Note insertion ratio	Notes inserted incorrectly / Total # of notes
S2-5	Pitch 1	Mean of difference between played pitch and score (mean, st. dev., min, max)
S6-9	Pitch 2	St. dev. of difference between playe1d pitch and score (mean, st. dev., min, max)
S10-13	Pitch 3	% of values deviating from score more than 1 st. dev. (mean, st. dev., min, max)
S14	DTW 1	Cost of DTW alignment
S15	DTW 2	Average deviation of alignment path from straight line
S16-21	Rhythm	Crest, bin resolution, skewness, kurtosis, roll-off, and power ratio of IOI histogram
S22	Note deletion ratio	Notes removed incorrectly / Total # of notes

found using the spearman correlation. Note that these descriptors are used in machine learning models which are able to parameterize the assessments as linear or non-linear combinations of the descriptors and hence there may or may not be a direct monotonic relationship between them.

We compute the spearman correlation coefficient of each of the 46 descriptors with each of the 4 human assessments and use the results in further experiments.

Descriptor Selection

In this experiment, we aim to identify the set of descriptors that are best able to explain the variance in each of the human assessments. This is computed by constructing linear regression models using various combinations or subsets of the descriptors at hand. Since the search space for the subsets is very large we first narrow down the search space by applying the following two criteria:

- Top 10 descriptors based on spearman correlation
- |Spearman correlation| > 0.25

We report the adjusted R-squared for the regression models and compare it to a model trained using all the descriptors.

In addition to this, we perform a sequential forward selection of the descriptors. In this experiment, we start with the best descriptor (the one that achieves highest R-squared) and iteratively check for the combination of descriptors with the highest adjusted R-squared and add it to our set of descriptors until the adjusted R-squared stops increasing.

Note that we do not perform validation using methods such as 10-fold cross-validation as used in previous work since our goal here is to understand how well each descriptor explains the variance in the entire dataset. Based on the identified descriptors, predictive models can be trained using a cross-validation scheme. However, the evaluation of such models is out of the scope of this study.

V. Results

The results for feature selection based on spearman correlation coefficient of each feature with the different human assessments are shown in Table 3. Each of the correlation values were statistically significant with $p < 10^{-4}$.

Table 3. Subsets of descriptors chosen based on spearman correlation coefficient of individual descriptors with the human assessments. The descriptors are arranged in decreasing order of correlation. Underlined feature indices indicate positive correlation.

	Top 10	|r| > 0.25
Musicality	N20, S17, S22, S10, N3, N15, N17, N2, S7, S1	N20, S17, S22, S10, N3, N15
Note Accuracy	S17, S14, S7, S2, S8, S20, N20, S6, S3, S10	S17, S14, S7, S2, S8, S20, N20
Rhythmic Accuracy	S17, N20, S14, S10, S3, S2, S20, S8, N15, S4	S17, N20, S14, S10, S3, S2, S20, S8, N15
Tone Quality	S17, N20, S1, S10, S2, S14, S3, S8, S7, S5	S17, N20, S1, S10, S2, S14, S3, S8, S7, S5

We can make the following observations:

- Most of the descriptors used are negatively correlated with the assessments. This makes sense because these descriptors are trying to summarize the mistakes made by the student performer.
- Descriptor S17 ranks very high for all assessment categories. The average spearman correlation with the assessments is -0.44. This feature is a measure of the IOI histogram's bin resolution. A high value indicates larger variance in tempo of the performance which is undesirable for the technical etude.
- Most of the top ranked descriptors are score-dependent descriptors. This is likely due to the descriptors being dependent on correct note segmentation of the performance. The additional score information allows, as expected, for a more robust note description and thus a more accurate extraction of performance parameters.

In Figure 1 we compare different regression models trained using 3 different subsets of descriptors: All descriptors, descriptors with correlation > 0.25 and the top 10 descriptors based on their correlation with the assessment. We observe that the smaller subsets are able to account for a large degree of variance explained by the entire set for Musicality, in particular. This is not true for Rhythmic Accuracy, which might be since the best descriptors (S17 and N20) are highly correlated and only 3 of the 12 Rhythmic Accuracy descriptors appear in the top 10. We also observe that the descriptors are poor at predicting Tone Quality. This is most likely due to the fact that we do not have descriptors for timbral characteristics of the performance.

Finally, Figure 2 shows the results for the sequential forward selection. We observe that after around 20 to 30 iterations over the descriptors, the models stop improving. This could be due to the fact that the remaining descriptors are not adding any new information and are not causing any improvement in the models' predictive accuracy. This calls for removal of redundant descriptors and addition of new descriptors. Another possible explanation for this is the curse of dimensionality (Friedman, 1997). This implies that given the number of descriptors, the amount of data is insufficient for the model to take advantage of additional information.

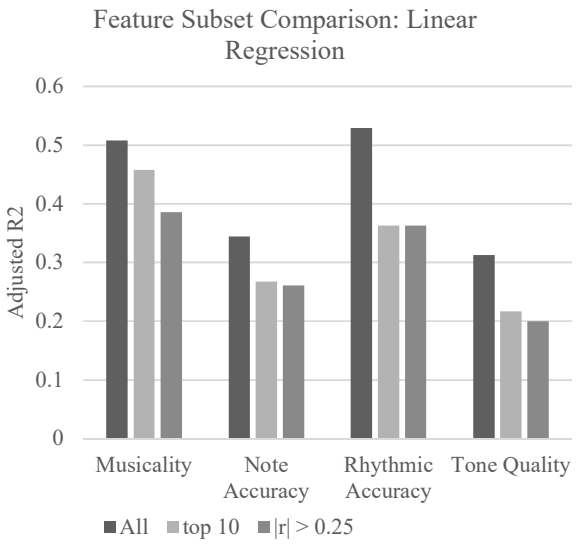

Figure 1. Comparison of linear regression models fitted with different subsets of descriptors.

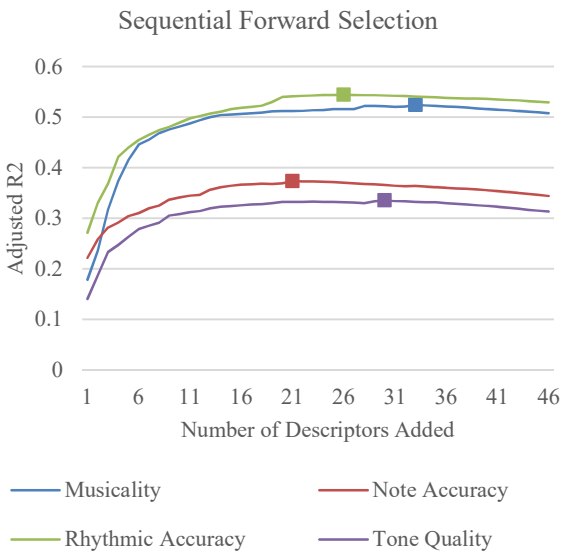

Figure 2. Learning curves for linear regression models trained using sequential forward selection. The boxes indicate the point where the adjusted R-squared starts decreasing.

From this experiment, we were able to identify the set of features that are able to explain the maximum variance in each assessment criteria. These descriptors are the ones that are selected for the model indicated by the box in Figure 2. We can also observe that these descriptors explain the variance to a greater degree than all the subsets of descriptors used in Figure 1. This is due to the fact that the descriptors chosen using correlation as the metric may be correlated amongst themselves while in the forward selection procedure, descriptors are added to the subset based on the increase in R-squared, leading to descriptors that are not highly correlated amongst themselves.

The first iteration of the experiment selects the descriptor that is best able to explain the variance among all descriptors. For Musicality, it is descriptor N20 which is the score-independent IOI histogram bin resolution (adjusted R2 = 0.18). For Note Accuracy, it is descriptor S14 which is the DTW cost (adjusted R2 = 0.22). For Rhythmic Accuracy, it is descriptor S17 which is the score-dependent IOI histogram bin resolution (adjusted R2 = 0.27). For Tone Quality it is descriptor S1 which is the note insertion ratio (adjusted R2 = 0.14).

We observe from Figure 2 that the performance for Musicality increases rapidly with the iterations. This is possibly due to the fact that Musicality is loosely defined, and the assessments are better explained with a combination of different kinds of descriptors. For Note Accuracy, the curve is flatter implying that the variance is captured in early iterations and subsequent descriptors are redundant. The DTW cost is a relevant descriptor for note accuracy since it is computed by aligning the pitch contour and the note sequence. For Rhythmic Accuracy, the best descriptor explains the variance to a greater degree than the other categories which may be attributed to the relevance of IOI histogram bin resolution. In the results for correlation analysis, it was shown to be among the top descriptors. In the case of Tone Quality, the descriptors perform the worst. This reiterates the fact that the current set of descriptors lack in terms of capturing timbral characteristics of the student music performance.

Conclusion

In this paper, we perform an in-depth analysis of 46 hand-crafted descriptors for the assessment of student alto saxophone performances to quantify their relevance. Our experiments show that a subset of the descriptors is well correlated with the human assessments. In addition, we find that the score-dependent descriptors are better correlated with the assessments compared to the score-independent descriptors.

We select subsets of descriptors with relatively high correlations with the assessments and construct linear regression models. We use the adjusted R-squared metric to compare the descriptor subsets with the whole set. The results from this experiment show that some of the top descriptors are able to account for a large degree of the variance explained by the entire set. The experiment for sequential forward selection shows that only around 30 of the 46 descriptors are selected for the models with the highest adjusted R-squared for each individual assessment category. This may be explained by redundancy in the descriptors or that the dimensionality is too high after a point and more data is required for improvement in regression performance.

As future work, we aim to remove the redundant descriptors and add new descriptors which can help capture a higher degree of variance in each assessment category focusing on adding features relevant to timbral characteristics.

Acknowledgements. We would like to thank the Florida Bandmasters Association (FBA) for kindly providing us with the dataset used in this work.

References

Abeßer, J., Hasselhorn, J., Dittmar, C., Lehmann, A., & Grollmisch, S. (2013, October). Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR), Marseille, France* (pp. 975–988)

Devaney, J., Mandel, M. I., & Fujinaga, I. (2012, October). A Study of Intonation in Three-Part Singing using the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT). In *Proceedings. of the International Society for Music Information Retrieval Conference (ISMIR), Porto, Portugal* (pp. 511-516).

Dittmar, C., Cano, E., Abeßer, J., & Grollmisch, S. (2012). Music information retrieval meets music education. In *Dagstuhl Follow-Ups* (Vol. 3). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, *1*(1), 55-77.

Han, Y., & Lee, K. (2014, October). Hierarchical Approach to Detect Common Mistakes of Beginner Flute Players. In *Proceedings. of the International Society for Music Information Retrieval Conference (ISMIR), Taipei, Taiwan* (pp. 77-82).

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons.

Mayor, O., Bonada, J., & Loscos, A. (2009, February). Performance analysis and scoring of the singing voice. In *Proceedings 35th AES International. Conference., London, UK* (pp. 1-7).

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, 69-84.

Palmer, C. (1997). Music performance. *Annual review of psychology, 48(1)*, 115-138.

Pati, K. A., Gururani, S., & Lerch, A. (2018). Assessment of Student Music Performances Using Deep Neural Networks. *Applied Sciences*, *8*(4), 507.

Vidwans, A., Gururani, S., Wu, C. W., Subramanian, V., Swaminathan, R. V., & Lerch, A. (2017, June). Objective descriptors for the assessment of student music performances. In *Proceedings of 2017 Audio Engineering Society (AES) Conference on Semantic Audio.*

Wesolowski, B. C. (2012). Understanding and developing rubrics for music performance assessment. *Music Educators Journal*, *98*(3), 36-42.

Wu, C. W., Gururani, S., Laguna, C., Pati, A., Vidwans, A., & Lerch, A. (2016, July). Towards the Objective Assessment of Music Performances. In *Proc. of the International Conference on Music Perception and Cognition (ICMPC)* (pp. 99-103).

Wu, C. W., & Lerch, A. (2018, February). Learned Features for the Assessment of Percussive Music Performances. In *Proceedings of the International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA* (Vol. 31).