Towards the Objective Assessment of Music Performances

Chih-Wei Wu, Siddharth Gururani, Christopher Laguna, Ashis Pati, Amruta Vidwans, Alexander Lerch

Center for Music Technology, Georgia Institute of Technology, USA
{cwu307, siddgururani, claguna3, ashis.pati, avidwans, alexander.lerch}@gatech.edu

ABSTRACT

The qualitative assessment of music performances is a task that is influenced by technical correctness, deviations from established performance standards, and aesthetic judgment. Despite its inherently subjective nature, a quantitative overall assessment is often desired, as exemplified by US all-state auditions or other competitions. A model that automatically generates assessments from the audio data would allow for objective assessments and enable musically intelligent computer-assisted practice sessions for students learning an instrument. While existing systems are already able to provide similar basic functionality, they rely on the musical score as prior knowledge. In this paper, we present a score-independent system for assessing student instrument performances based on audio recordings. This system aims to characterize the performance with both well-established and custom-designed audio features, model expert assessments of student performances, and predict the assessment of unknown audio recordings. The results imply the viability of modeling human assessment with score-independent audio features. Results could lead towards more general software music tutoring systems that do not require score information for the assessment of student music performances.

I. INTRODUCTION

Music performance, according to Palmer, is one of the most complex serial actions produced by human beings, requiring the interpretation of musical ideas, the planning of the retrieved musical units, and the transformation of these thoughts into movements (Palmer, 1997). The qualitative assessment of performances by teachers and peers is, despite its inherent difficulty due to the subjective nature of the task, an essential part of music education. The teacher has to provide structured quality assessment, possibly quantifying different aspects of a performance and thus providing systematic feedback in order to facilitate improvement and reach the learning outcomes. However, as pointed out by Thompson and Williamon (Thompson & Williamon, 2003), the bias of the evaluators and the highly correlated categories in the structured assessment could impact the discriminability and the fairness of this approach. A computational approach that models the human cognition of the music performances and provides consistent and reproducible feedback might be a potential solution to this issue. It might also be used as a tool to provide feedback to students during practice sessions without an instructor.

Approaches and tools from the research field Music Information Retrieval (MIR) are being utilized more and more in software solutions for music education (Dittmar, Cano, Abeßer, & Grollmisch, 2012). With the advancement in research topics such as source separation (Huang, Kim, Hasegawa-Johnson, & Smaragdis, 2014) and music transcription (Benetos, Dixon, Giannoulis, Kirchhoff, & Klapuri, 2013), different music learning systems with reliable functionalities can be created, offering objective and repeatable evaluation to the users. Commercial software such as SmartMusic (http://www.smartmusic.com Last access: 2016/04/24) and Yousician (https://get.yousician.com Last access: 2016/04/24) are examples of such systems.

Since most of these systems require the score as additional input, their applications are limited and frequently depend on proprietary curated content provided by the manufacturer. In this paper, we explore the idea of building a score-independent system. A set of well-established audio features used in MIR tasks (Tzanetakis & Cook, 2002) is compared to a set of custom-designed features in a machine learning based regression model predicting the assessments of human experts. The goal is to investigate whether score-independent descriptors can be meaningful for the general assessment of student music performances. This paper is structured as follows: Sect. 2 introduces the related work on music performance assessment. The details of the dataset used in this work are described in Sect. 3. The methodology and the experimental setups are mentioned in Sect. 4. Finally, the experiment results and conclusions are presented in Sect. 5 and 6, respectively.

II. RELATED WORK

Music performance analysis is a research field that involves the observation, extraction, and modeling of important parameters in music performances. Early research focused on the analysis of symbolic data collected from external sensors or MIDI devices. More recently, the focus has gradually shifted to the analysis of audio recordings. For example, Abeβer et al. proposed a system that automatically assesses the quality of vocal and instrumental performances of 9th and 10th graders (Abeßer, Hasselhorn, Dittmar, Lehmann, & Grollmisch, 2013). The assessment is obtained by modeling the relationship between score-based features and the experts' ratings. The rating ranges from 1 to 4: 1 being the best, and 4 being the worst performance quality. A four-class classifier is trained to assess the performance, and the evaluation results show that the system is able to classify the performances although exhibiting some confusion between adjacent ratings. Another example is the score-informed piano tutoring system presented by Fukuda et al. (Fukuda, Ikemiya, Itoyama, & Yoshii, 2015), which applies automatic music transcription and audio to score alignment to detect the mistakes in the user's performance. The system also includes a score-simplification functionality to motivate the users by reducing the difficulty of a given score. The evaluation results show that the system can transcribe the audio input with high accuracy, and highlight the mismatches between the score and the performance with occasional octave errors.

In the studies mentioned above, a score is usually a prerequisite for the automatic assessment. However, in certain use cases, such as free practice or improvisation, these systems are not directly applicable since no score is available.

Nakano presented an automatic system that evaluates the singing skill of the users (Nakano, Goto, & Hiraga, 2006). The system is trained based on the extracted pitch interval accuracy and vibrato features without any score input. The evaluation results show that the system is able to classify the performance into two classes (good or poor) with high accuracy. Mion and De Poli proposed a system that classifies music expressions based on score-independent audio features (Mion & De Poli, 2008). With instantaneous and event-based features such as spectral centroid, residual energy, and notes per second, the system is able to recognize four different musical expressions of violin, flute, and guitar performances. These examples show the potential of analyzing a recorded music performance without the need for the underlying score.

III. DATASET

The dataset used for this study is kindly provided by the Florida Bandmasters Association (FBA). The dataset contains 3344 audio recordings of the 2013–2014 Florida all-state auditions with accompanying expert assessments. The participating students are divided into three groups, namely middle school (7^{th} and 8^{th} grade), concert band (9^{th} and 10^{th} grade), and symphonic band (11^{th} and 12^{th} grade). A total number of 19 types of instruments are played during the auditions. More details are shown in Table 1. All of the auditions are recorded at a sampling rate of 44100 Hz and are encoded with MPEG-1 Layer 3.

For pitched instruments, each audition session includes 5 different exercises, which are lyrical etude, technical etude, chromatic scale, 12 major scales, and sight-reading. For percussion instruments, the audition session also includes 5 different exercises, which are mallet etude, snare etude, chromatic scale (xylophone), 12 major scales (xylophone), and sight-reading (snare). Each exercise is graded by human experts with respect to different assessment categories, such as musicality, note accuracy, rhythmic accuracy, tone quality, artistry, and articulation, etc. In our experiments, all of the ratings are normalized to a range between 0 and 1, with 0 being the minimum and 1 being the maximum score.

To narrow the scope of this study, only a small subset of this original dataset is used. This subset includes the recordings of one exercise from one pitched instrument (alto sax) and one percussion instrument (snare drum) performed by middle school students. These two instruments have been selected because they have relatively higher number of recordings in the dataset (122 and 98, respectively).

Table 1. Statistics of the 2013-2014 FBA audition dataset

	# Files	Total Duration (mins)	Avg. Duration (mins)
Middle School	1099	5460	4.96
Concert Band	1046	5280	5.04
Symphonic Band	1199	6540	5.45
List of All Instruments			
Alto Sax, Baritone Sax, Bass Clarinet, Bass Trombone, Bassoon, Bb Clarinet, Bb Contrabass Clarinet, Eb Clarinet, English Horn, Euphonium, Flute, French Horn, Oboe, Percussion, Piccolo, Tenor Sax, Trombone, Trumpet, Tuba			

IV. EXPERIMENTS

A. Feature Extraction

To represent the audio signals in the feature space, two types of features are extracted: 1) baseline features and 2) designed features.

The baseline features (d = 17, d is the dimensionality) are computed block-by-block with a window size of 1024 samples and a hop size of 256 samples. A Hann window is applied to each block. The baseline features include: spectral centroid, spectral rolloff, spectral flux, zero-crossing rate, and 13 Mel Frequency Cepstral Coefficients (MFCCs). The features are implemented according to the definitions in (Lerch, 2012). To represent each recording with one feature vector, a two-stage feature aggregation process has been applied. In the first stage, the block-wise features within a 250ms texture window are first aggregated by their mean and standard deviation. In the second stage, all of these meta-features are aggregated again into one single vector with their mean and standard deviation, resulting in one baseline feature vector (d = 68) per recording.

The designed features for the percussion instruments are used to capture the rhythmic aspects of the audio signal. The resulting features (d = 18) per recording are shown as follows:

- Inter-Onset-Interval (IOI) histogram statistics (d = 7): This set of features is designed to describe the rhythmic characteristics of the played onsets. An IOI histogram is computed with 50 bins. Next, the standard statistical measures crest, skewness, kurtosis, rolloff, flatness, tonal power ratio, and the histogram resolution are extracted.
- Amplitude histogram statistics (d = 11): This set of features is designed to capture the amplitude variations of the played onsets. The amplitude of the waveform is first converted into dB, and the amplitude histogram is calculated with 50 bins. Next, the same statistical measures as above are extracted. Additionally, the length of the exercise, the standard deviation of the amplitude in both linear and dB scale, and the Root Mean Square (RMS) of the entire waveform are included as features.

The designed features for the pitched instruments are intended to describe various dimensions of the performances, namely the pitch, dynamics, and timing characteristics. Most of these features are extracted at the note-level after a simple segmentation process based on the quantized pitch contour. The pitches are detected using an autocorrelation function based pitch tracker. The note-level features are:

- Note steadiness (d = 2): These two features are designed to find fluctuations in the pitch of a note. For each note, the standard deviation of pitch values and the percentage of values deviating more than one standard deviation from the mean are computed.

- Amplitude deviation (d = 1): This feature aims to find the uniformity of the amplitude of a note. For each note, the standard deviation of the RMS is computed.
- Amplitude envelope spikes (d = 1): This feature describes the spikiness of the note amplitude over time. The number of local maxima of the smoothed derivative of the RMS is computed per note.

Once the above features are extracted, their mean, maximum, minimum, and standard deviation across all the notes are calculated to represent the recording. In addition, the following exercise-level features are extracted:

- Average pitch accuracy (d = 1): This feature shows the consistency of the notes played. The histogram of the pitch deviation from the closest equally tempered pitch is extracted with a 10 cent resolution. The area under the window (width: 30 cent) centered around the highest peak is considered as the feature.
- Percentage of correct notes (d = 1): For this feature, each note is labeled either correct or incorrect, and the percentage of correct notes across the entire exercise is computed as the feature. A note is labeled correct if the percentage of pitch values with a deviation from the mean pitch is lower than a pre-defined threshold.
- Timing accuracy (d = 7): These features are computed from the IOI histogram of the note onsets. Note onsets are computed from the pitch contour. The features used are the same as the features that were used for percussive instruments.

The resulting feature vector for the designed features for pitched instruments has the dimension d = 25.

B. Feature Extraction

Using the extracted features from the audio signals, a Support Vector Regression (SVR) model with a linear kernel function is trained to predict the human expert ratings. The libsvm (Chang & Lin, 2011) implementation of this model is used with default parameter settings. A Leave One Out cross-validation scheme is adopted to train and evaluate the models.

C. Experimental Setup

To compare the effectiveness of the baseline features versus the designed features, two sets of experiments are conducted. In the first set of experiments, the baseline and designed features are extracted from the pitched instrument recordings and used to build the regression models for predicting labels such as artistry, musicality, note accuracy, rhythmic accuracy, and tone quality. In the second set of experiments, the same procedure is used to predict musicality, note accuracy, and rhythmic accuracy for the percussion instrument recordings. An outlier removal process is included for all the experiments. This process removes the training data with the highest prediction residual (prediction minus actual rating) and is repeated until 5% of the dataset is eliminated. By removing the outliers, the regression models should be able to better capture the underlying patterns in the data.

D. Evaluation Metrics

The performances of the models are evaluated by the following standard statistical metrics: the correlation coefficient r and its corresponding p-value, the R^2 value, and the standard error. These metrics are typically used to measure the strength of the regression relationship between the predictions and the actual ratings. More details of the mathematical formulations can be found in (McClave & Sincich, 2003).

V. RESULTS AND DISCUSSION

The results of the first and second set of experiments are shown in Table 2. Most of the entries have $p \ll 0.05$ except for artistry in the baseline features and note accuracy in both features of the pitched instrument; therefore, the p-values are not shown in the table.

The following trends can be observed: first, the baseline feature set is not able to capture enough meaningful information to create a usable model.

Second, compared to the baseline features, the designed features lead to general improvements in all the metrics for both pitched and percussion instruments. In most cases, the R^2 values increase by at least 30%, illustrating the effectiveness of the designed features at characterizing the music performances.

Third, musicality is the assessment that is modeled best for both pitched and percussion instruments. The highest correlation coefficient and R^2, 0.7307 and 0.5254 respectively, are achieved when using designed features to predict musicality for pitched instrument recordings. This could be explained by the fact that musicality is a relatively abstract description that covers most aspects of a performance, and therefore, it is likely to be related to the general quality of the performance, making it relatively easy to model. Table 3 shows the result of an inter-label investigation, correlating one type of rating with the others. It is found that musicality tends to be highly correlated with the other assessment categories. This result further confirms the relationship between musicality and overall performance quality.

Fourth, note accuracy for the pitched instrument is the worst performing category. Compared to a percussion instrument, more information might be required to model the note accuracy of a pitched instrument properly. The proposed features might be unsuitable for capturing the relevant information. This result could imply the necessity of including score information in order to improve the model of certain categories.

Last but not least, the baseline features did not perform better than the designed features in predicting the tone quality. This seems to contradict the intuition that timbre features are directly related to the tone quality. On the one hand, this could mean that other, not extracted low-level characteristics such as the initial attack phase of each note might have larger impact on the tone quality than the spectral envelope modeled by our features. On the other hand, this result might suggest a connection between the perceived tone quality and the high-level information captured by the designed features, such as note steadiness and pitch accuracy.

Table 2. Experiment results of the regression models

Inst	Feature	Label	r	R^2	Std. Err
Pitched (Alto Sax)	Baseline	Artistry	0.1559	-0.6750	0.2025
		Musicality	0.4698	-0.0693	0.0914
		Note Acc.	0.1134	-0.5735	0.1765
		Rhyt. Acc.	0.4099	-0.0783	0.1674
		Tone Qual.	0.3659	-0.2372	0.1056
	Designed	Artistry	0.4548	0.1635	0.1385
		Musicality	0.7307	0.5254	0.0627
		Note Acc.	0.1578	-0.0840	0.1465
		Rhyt. Acc.	0.6252	0.3727	0.1195
		Tone Qual.	0.5249	0.2350	0.0810
Percussion (Snare)	Baseline	Musicality	0.4537	0.0661	0.1709
		Note Acc.	0.4853	0.0695	0.1592
		Rhyt. Acc.	0.3835	-0.0025	0.1685
	Designed	Musicality	0.6489	0.4174	0.1293
		Note Acc.	0.5674	0.2393	0.1384
		Rhyt. Acc.	0.5467	0.2745	0.1497

Table 3. Inter-label cross-correlation results for musicality

Pitched Instrument					
	Artistry	Musicality	Note	Rhythmic	Tone
Musicality	0.7352	1	0.6439	0.7967	0.7037
Percussion Instrument					
Musicality	N/A	1	0.6790	0.8090	N/A

VI. CONCLUSION

In this paper, a system that automatically assesses student music performances based on score-independent audio features is presented. With the presented features the system is, to a certain degree, able to model and predict the ratings given by human experts without prior knowledge of the underlying scores. The results of the experiments show that, for both pitched and percussion instruments, the designed features perform better than the baseline features. Overall, the results imply the general feasibility of using score-independent features in assessing student music performances, although the presented features only capture part of the performance characteristics to be modeled.

The challenges and future directions of this research are: first, the current dataset, in spite of being diverse, only contains a few samples for specific combinations of instrument and group. To build a generic model, more data is needed. Second, although the designed features are shown to be useful, they require a lot of domain knowledge and might not be directly applicable to other datasets. An alternative solution is to apply a feature learning method such as Sparse Coding (Abdallah & Plumbley, 2006), which could automatically learn relevant features and potentially achieve a higher performance. Finally, to further investigate the necessity of the musical score especially in certain assessment categories such as note accuracy, score-based features should be explored and compared in the future.

ACKNOWLEDGMENT

The authors would like to thank the Florida Bandmasters Association for providing the dataset used in this study.

REFERENCES

Abdallah, S. A., & Plumbley, M. D. (2006). Unsupervised analysis of polyphonic music using sparse coding. *IEEE Transactions on Neural Networks*, *17*(1), 179–196.

Abeßer, J., Hasselhorn, J., Dittmar, C., Lehmann, A., & Grollmisch, S. (2013). 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)* (pp. 975–988).

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. (2013). Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27:1–27:27.

Dittmar, C., Cano, E., Abeßer, J., & Grollmisch, S. (2012). Music Information Retrieval Meets Music Education. In *Multimodal Music Processing* (Vol. 3, pp. 95–120). Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik.

Fukuda, T., Ikemiya, Y., Itoyama, K., & Yoshii, K. (2015). A Score-informed Piano Tutuoring System with Mistake Detection and Score Simplification. In *Proceedings of Sound and Music Computing Conference (SMC)*.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014). Singing-voice separation from monaural recordings using deep recurrent neural networks. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*.

Lerch, A. (2012). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. John Wiley & Sons.

McClave, J. T., & Sincich, T. (2003). *Statistics* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Mion, L., & De Poli, G. (2008). Score-Independent Audio Features for Description of Music Expression. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 458–466.

Nakano, T., Goto, M., & Hiraga, Y. (2006). An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)* (pp. 1706–1709).

Palmer, C. (1997). Music performance. *Annual Review of Psychology*, *48*, 115–138. http://doi.org/10.1146/annurev.psych.48.1.115

Repp, B. H. (1996). Patterns of note onset asynchronies in expressive piano performance. *The Journal of the Acoustical Society of America*, *100*(6), 3917–3932.

Shaffer, L. (1984). Timing in solo and duet piano performances. *The Quarterly Journal of Experimental Psychology*, *36*(4), 577–595.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception*, *21*(1), 21–41.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, *10*(5), 293–302.