Evaluation of Deep Neural Networks for Musical Performance Assessment

Pavan Seshadri [pseshadri9@gatech.edu]

Abstract

Music performance assessment is a very subjective and intensive task, which currently requires expert human analysis of a performance to give feedback on technical and aesthetic parts of a performance. There has been recent research in using Deep Neural Networks to learn representations of an audio recording of a performance to accurately score a performance on a variety of criteria consistent with an expert judge; however, this research has not explored the ability of these networks to generalize to instruments outside of its training set. This project aims to evaluate the performance of Deep Neural Networks on instruments outside of its training set. Overall, with current data limitations, it is shown that current deep learning-based techniques do not generalize well when tested on instruments outside of its training set. The results serve to highlight current limitations within this field of research.

1. Background

There is a need to make music education more accessible through the use artificially intelligent software tutoring systems. These systems allow for individualized progress and feedback while providing accessibility to students across all backgrounds. A first step towards this goal is to build systems that allow reproducible and objective analysis results for the inherently subjective task of music performance assessment. While generally experienced professionals judge various technical and aesthetic aspects of the performance, there has been research focused on building computational models to analyze audio recordings of a student music performance and rate it along various criteria. I plan to evaluate the performance of Deep Neural Networks (DNNs) for this task, as previous research in this area has shown that DNNs have outperformed baseline models using hand crafted features designed to extract relevant aspects of a performance [1]. Performers are tasked with interpreting the directions from a musical score and translating it to an acoustic rendition. This includes modifying various performance parameters such as tempo and timing, dynamics, intonation, and tone quality in order to craft a unique performance [2].

1.1 Related Work.

Music Performance Analysis aims to understand and model the impacts of these deviations from the baseline score on a human listener [3]. Early research centered around analyzing symbolic data extracted from MIDI devices [4,5]. Recent research has begun to focus on analyzing raw audio [3,6]. In human assessment, music instructors must discern the individual subjective qualities and give a holistic score. The idea of what is “good” or “bad” is not well defined, so the ratings of music instructors often have high variance [7,8]. Most attempts in automatic music assessment systems have involved extracting hand crafted features of an audio signal and feeding to a classifier algorithm to judge the quality of the performance [9-13]. This approach relies on knowledge of experts to extract relevant features to classify. I aim to evaluate supervised learning-based methods for this task of Music Performance Assessment. I specifically aim to compare performance of DNNs trained on multiple instruments on one single instrument. Gururani et. Al [1] compared performance of DNNs trained on a dataset comprised of three different instruments and tested on a similarly mixed dataset. I aim to expand this to evaluate DNN performance when trained on one single instrument and tested on one single instrument, including instruments outside of the training set. This project aims to progress the use of supervised learning in MPA.

2. Methods

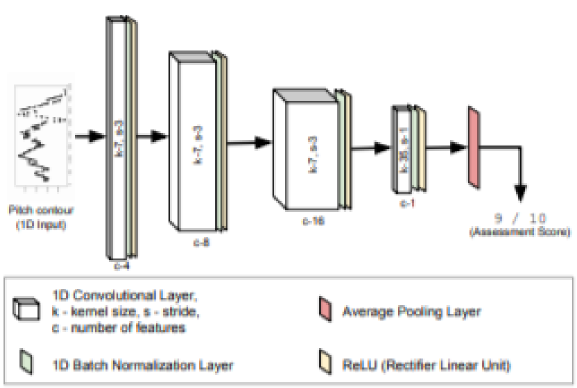
This project focuses on evaluating DNN-based regression models that can predict ratings given by expert human judges for pitched wind instruments. I experiment specifically with pitch contour representations, and Mel-spectrogram representations, for both a high- and low-level encoding of the audio data, respectively. The pitch contour representations of the audio were computed for each individual raw audio file and were precomputed within the used dataset. Mel Spectrogram representations were computed for each datapoint just before training and testing the models. Assessment ratings are predicted on the following metrics, musicality, note accuracy, and rhythmic accuracy. Mel Spectrogram input models also were tasked with judging tonality. Musicality judges the expressiveness of the performance in qualities such as dynamics and articulation. Note accuracy and rhythmic accuracy judge the difference of played notes and rhythms, respectively, to that of the written score. Tonality judges the tone quality of the instrument in the performance.

2.1 Dataset

The dataset that the models are trained on are audio recordings and ratings of auditions from the Florida Bandmaster's Association (FBA) from 2013 to 2018. This dataset contains raw audio recordings, and pitch contour representations from each audio recording from three different levels of audition, Middle School, Concert Band, and Symphonic Band. Provided recordings include auditions from alto saxophone, Bb clarinet, and flute players. Each recording contains several exercises which vary by instrument, level, and year. I specifically used the Middle School level data, and built separate datasets for each individual instrument, alto saxophone, flute, and clarinet. The datasets contained 927 Bb Clarinet recordings, 993 Flute recordings, and 696 Alto Saxophone recordings. Scores for each category are in the range [0, 1].

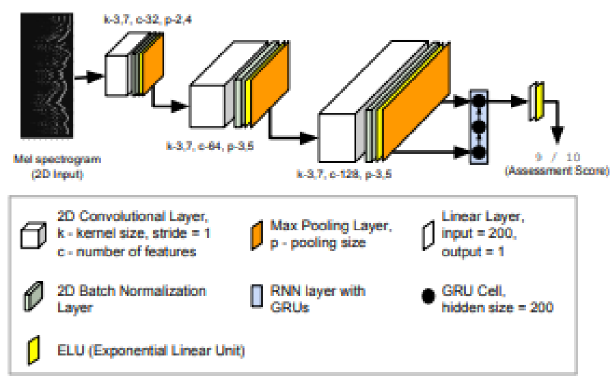
2.2 Pitch Contour input Model Architecture

A Fully Convolutional Network was used for this regression task, using pitch contours representations as input. The following model architecture adapted from Gururani et. Al [1]. was used to predict ratings.



2.3 Mel Spectrogram input Model Architecture

A Convolutional Recurrent Model was used for this regression task, using Mel Spectrogram representations as input. The following model architecture adapted from Gururani et. Al [1]. was used to predict ratings.



2.3 Experimentation Procedure and Implementation

Each model was trained on one instrument and output one evaluation metric. This provided a total of 9 models to evaluate for the pitch contour model, with 3 evaluation metrics per 3 separate instruments, and 12 models to evaluate for the Mel-spectrogram model, with 4 evaluation metrics per 3 instruments. Training, testing, and validation were split from the overall dataset using an 80/10/10 split. The data was also shuffled before the split to reduce the potential of distribution differences between each set. The models are judged by a computed coefficient of determination (r^2 score) based on the ground truth of a testing set, and a set of predictions derived from the aforementioned set. These scores were compared to a baseline evaluation using a model trained on all three instruments. All implementation, experimentation, and analysis was implemented using the Python programming language, using the PyTorch framework to implement and train models. Data visualization was implemented using the matplotlib framework. Model implementation, training, and evaluation code was altered and adapted from code provided from the paper from Gururani et. Al [1].

2.3 Hyperparameters

Mean Squared Error was used as the loss function, with stochastic gradient descent as the optimizer using parameters, 1e-2 as the learning rate, 1e-5 as the weight decay, and 0.9 as the momentum. The models were trained over 2000 epochs, with a mechanism to halt training if performance over the validation set has not improved over the past 200 epochs. It was found that a higher learning rate positively affected performance and reduced training epochs necessary, likely due to a reduced size dataset. Stochastic Gradient Descent was also found to marginally improve performance over the Adam optimizer algorithm.

3. Results

As a baseline, one model was trained on a dataset of all the instruments, for each assessment metric, and tested on each individual instrument, using the pitch contour model. The results are as follows. Each number corresponds to the r^2 score of the test set and its predictions.

Instrument	Musicality	Note Accuracy	Rhythm Accuracy
Clarinet	0.637	0.613	0.636
Flute	0.436	0.415	0.349
Sax	0.067	0.174	0.399

From the beginning, we see poor performance on the saxophone dataset for musicality and note accuracy, very high performance on clarinet data across the board and average performance on the flutes overall. Previous results from Gururani et. Al [1] indicate about 0.52, 0.43, and 0.35 as previous baselines for musicality, note accuracy, and rhythm accuracy, respectively. Initially this potentially highlights inconsistencies within the dataset, as a model trained on all instruments, with similar numbers of datapoints per instrument, performs drastically different on different sets containing different instruments.

3.1. Saxophone Results

Results for the saxophone-trained models for the pitch contour model are as follows.

Instrument	Musicality	Note Accuracy	Rhythm Accuracy
Clarinet	-0.29	0.28	0.396
Flute	-0.705	-0.095	-0.839
Sax	**0.41**	**0.2**	**0.15**

The results for the saxophone-trained models for the Mel-Spectrogram model are as follows.

Instrument	Musicality	Note Accuracy	Rhythm Accuracy	Tonality
Flute	-0.26	-1.5	-0.83	-0.65
Clarinet	-0.12	-1.7	-0.83	-0.8
Sax	**.31**	**0.2**	**0.47**	**.16**

For the pitch contour model, we see mediocre results of the saxophone-trained models when tested on saxophone data, extremely poor performance when tested on flute, extremely poor musicality performance when tested on clarinet, and mediocre results for Note Accuracy and Rhythm Accuracy. The Mel-Spectrogram model performs similarly, except for much better, but still poor performance testing on clarinet for note accuracy and rhythm accuracy. Poor generalization for musicality is seen when the model learns on both pitch contours and mel-spectrograms, so neither a lack of low-level data, nor including it allowed the model to generalize. One possible explanation is inconsistencies within the dataset—As these scores are subjective in the first place, it is possible that the distribution and criteria of scores for saxophones is different than flute and clarinet, and a comparable performance between the two instruments may not receive the same score. The reduced datapoints for saxophones may exacerbate this as well.

3.2 Flute Results

Results for the flute-trained models for the pitch contour model are as follows.

Instrument	Musicality	Note Accuracy	Rhythm Accuracy
Clarinet	.15	.11	.48
Flute	**.33**	**.25**	**.505**
Sax	-0.67	-4.07	-0.84

Results for the flute-trained models for the Mel-Spectrogram model are as follows.

Instrument	Musicality	Note Accuracy	Rhythm Accuracy	Tonality
Flute	**.43**	**.47**	**.41**	**.31**
Clarinet	.53	-.87	-1.07	.21
Sax	.53	-.9	-.91	.22

Similarly, to the saxophone-trained models, we see mediocre to decent results when tested on the same-instrument test set for both models. Cross-instrument tests for the pitch contour model on the clarinet testing set provides poor results for musicality and note accuracy, but decent results for Rhythm Accuracy. The Mel-Spectrogram was largely the same, besides musicality, where it performed decently on both clarinet and saxophone, and relatively equal but worse for tonality on these instruments. The difference in performance between the cross-instrument tests for clarinet versus saxophone in the pitch contour models serve as additional evidence for potential dataset inconsistencies for the saxophone set.

3.3 Clarinet Results

Results for the clarinet-trained models for the pitch contour model are as follows.

Instrument	Musicality	Note Accuracy	Rhythm Accuracy
Clarinet	**.61**	**.59**	**.62**
Flute	0.307	0.294	0.152
Sax	-1.95	-0.703	-0.535

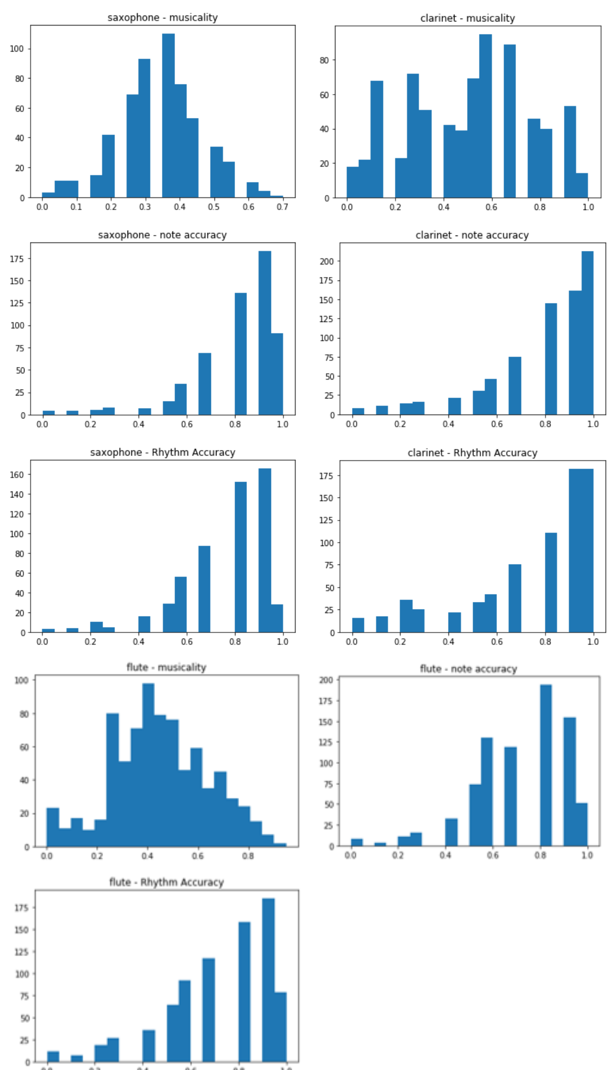
Results for the clarinet-trained models for the Mel-spectrogram model are as follows.

Instrument	Musicality	Note Accuracy	Rhythm Accuracy	Tonality
Flute	.37	-1.5	-1.34	.20
Clarinet	**.58**	**.55**	**.58**	**.45**
Sax	.28	-1.52	-1.42	.16

Again, similarly to the flute-trained trials, we see extremely poor results when tested on saxophone data for both inputs, and middling results when tested on flute for the pitch contour model, whereas the mel-spectrogram stays poor. However, mirroring the all-instruments trained model, the clarinet trained model performs exceptionally well, and consistent with the clarinet performance benchmark, when tested on clarinet data. Coupled with the context of same-instrument model performance on flute and saxophone, which for several categories also perform consistent with the baseline, it becomes increasingly plausible that dataset inconsistencies are present. There is evidence here that potentially, clarinet data is either more consistently scored by judges, or that the methods of pitch contour extraction may be more accurate on clarinet compared to flute and saxophone.

3.4 Dataset Distribution Comparisons

As shown by previous results, there is potentially high possibility of dataset inconsistencies between instruments, which made it necessary to compare. Histograms of the distributions of scores across each assessment metric for each instrument are as follows.



For musicality, we see that the saxophone follows a normal-like distribution, while the clarinet data is more uniform, and the flute data is less normal-like, with less data at the lower end of scores. Interestingly, for the other assessment metrics, all are similarly skewed left. This gives more legitimacy to assessment criteria differing across instruments. Given the subjectivity of the problem at hand, it is very possible that a similar score for each instrument do not correspond to a similar performance for each instrument. While lower datapoints is also likely a factor, it is possible that the grading criteria for saxophones is noticeably different than that of clarinet and flute.

Additional analysis into predictions vs. ground truth distributions showed that the models were predicting values very far from the mean of the ground truth distribution for many runs, causing negative r^2 coefficients. At times, the distribution shape itself was markedly different from prediction to ground truth, further exemplifying the poor performance for most cross-test runs.

3.4 Overall Conclusions

Overall, these models are unable to generalize across other instruments that are absent from the training set. Repeated trials reducing model capacity and further altering hyperparameters to combat overfitting was found unsuccessful in increasing performance in the cross-instrument tests, and often reduced performance in the same-instrument tests. Dataset labelling inconsistency potential is strengthened by these results and by the non-significant distribution differences for two of the three assessment criteria. The pitch contour and mel-spectrogram results staying relatively equal disprove the notion that one method may be encoding too much specific information to generalize to instruments well, as neither metric was able to perform well in cross-tests reliably across each instrument for any metric. Some level of overfitting can be expected when feature learning over a reduced variance dataset, as was the case with this project, by removing instrument variety and datapoints from the original training set from Gururani et. Al[1]; however no tested methods proved successful in curbing this overfitting to allow the models to generalize to instruments outside of the testing set consistently.

4. Future Work

Additional analysis into dataset inconsistencies is likely needed to properly deduce the reason behind the extremely poor generalization of these models. Different model architectures may need to be explored to better learn instrument-agnostic characteristics of a performance as well. Current methods are clearly overfitting to the instrument of the testing set and perform poorly when given an instrumental performance outside of its testing set.

5. Acknowledgements

This project was done under the guidance of Professor Alexander Lerch and PhD student Ashis Pati. Funding assistance was provided by a President's Undergraduate Research Award (PURA) grant.

References

[1] Pati, K.A.; Gururani, S.; Lerch, A. Assessment of Student Music Performances Using Deep Neural Networks. Appl. Sci. 2018, 8, 507..

[2] Clarke, E. Understanding the Psychology of Performance. In Musical Performance: A Guide to Understanding; Cambridge University Press: Cambridge, UK, 2002; pp. 59–72.

[3] Lerch, A. Software-Based Extraction of Objective Parameters From Music Performances. Ph.D. Thesis, Technical University of Berlin, Berlin, Germany, 2008.

[4] Palmer, C. Mapping musical thought to musical performance. J. Exp. Psychol. 1989, 15, 331.

[5] Repp, B.H. Patterns of note onset asynchronies in expressive piano performance. J. Acoust. Soc. Am. 1996, 100, 3917–3932.

[6] Dixon, S.; Goebl, W. Pinpointing the beat: Tapping to expressive performances. In Proceedings of the 7th International Conference on Music Perception and Cognition (ICMPC), Sydney, Australia, 17–21 July 2002; pp. 617–620

[7] Wesolowski, B.C.; Wind, S.A.; Engelhard, G. Examining rater precision in music performance assessment: An analysis of rating scale structure using the Multifaceted Rasch Partial Credit Model. Music Percept. 2016, 33, 662–678.

[8] Thompson, S.; Williamon, A. Evaluating evaluation: Musical performance assessment as a research tool. Music Percept. 2003, 21, 21–41

[9] Nakano, T.; Goto, M.; Hiraga, Y. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Pittsburgh, PA, USA, 17–21 September 2006; pp. 1706–1709.

[10] Knight, T.; Upham, F.; Fujinaga, I. The potential for automatic assessment of trumpet tone quality. In Proceedings of the International Society of Music Information Retrieval Conference (ISMIR), Miami, FL, USA, 24–18 October 2011; pp. 573–578.

[11] Dittmar, C.; Cano, E.; Abeßer, J.; Grollmisch, S. Music Information Retrieval Meets Music Education. In Multimodal Music Processing; Müller, M., Goto, M., Schedl, M., Eds.; Dagstuhl Follow-Ups Series; Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik: Wadern, Germany, 2012; Volume 3.

[12] Abeßer, J.; Hasselhorn, J.; Dittmar, C.; Lehmann, A.; Grollmisch, S. 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, 15–18 October 2013.

[13] Romani Picas, O.; Parra Rodriguez, H.; Dabiri, D.; Tokuda, H.; Hariya, W.; Oishi, K.; Serra, X. A Real-Time System for Measuring Sound Goodness in Instrumental Sounds. In Proceedings of the 138th Audio Engineering Society Convention, Warsaw, Poland, 7–10 May 2015.