SCORE-INFORMED NETWORKS FOR MUSIC PERFORMANCE ASSESSMENT

Jiawen Huang **Yun-Ning Hung** **Ashis Pati**
Siddharth Gururani **Alexander Lerch**
Center for Music Technology, Georgia Institute of Technology, USA
{jhuang448,amyhung,ashis.pati,siddgururani,alexander.lerch}@gatech.edu

ABSTRACT

The assessment of music performances in most cases takes into account the underlying musical score being performed. While there have been several automatic approaches for objective music performance assessment (MPA) based on extracted features from both the performance audio and the score, deep neural network-based methods incorporating score information into MPA models have not yet been investigated. In this paper, we introduce three different models capable of score-informed performance assessment. These are (i) a convolutional neural network that utilizes a simple time-series input comprising of aligned pitch contours and score, (ii) a joint embedding model which learns a joint latent space for pitch contours and scores, and (iii) a distance matrix-based convolutional neural network which utilizes patterns in the distance matrix between pitch contours and musical score to predict assessment ratings. Our results provide insights into the suitability of different architectures and input representations and demonstrate the benefits of score-informed models as compared to score-independent models.

1. INTRODUCTION

A performance is a sonic rendition of a written musical score (in the case of Western classical music). The characteristics of a music performance play a major role in how listeners perceive music, even if performances are based on the same underlying score [14]. To perform a musical piece, the performer must first parse the score, interpret or modify the musical information, and utilize complex motor skills to render the piece on their instrument [21].

From the perspective of the performer, mastery over the art of music performance is often a journey spanning several years of instruction and practice. A major factor in learning and improving one's skill as a performer is to analyze and obtain feedback regarding the performance. Due to the complex nature of music performance, students require regular feedback from trained professionals. Teachers are expected to grade or rate students based on various performance criteria such as note accuracy or musicality. These criteria are often ill-defined and subject to interpretation, thus making objective and consistent music performance assessment (MPA) rather difficult [26, 29]. Regardless, this subjective manner of MPA is still used, e.g., in school systems where ensemble members are selected based on instructors' assessments of student auditions.

Wu et al. discussed the notion of objective descriptors (features) which are potentially useful for automatic MPA [30]. Such features are computed by applying signal processing methods to recorded performances and are used to model teachers' assessments of the performances using machine learning. With the rise of deep learning, neural networks were found to outperform the classical pipeline of feature extraction followed by regression [22]. However, one issue with these approaches is that they ignore the score that the students are meant to play. We will refer to such approaches as *score-independent*. The idea of incorporating score-based features utilizing audio to score alignment was explored, e.g., by Vidwans et al. [27]. Further analysis of hand-crafted features for MPA showed the relative importance of score-based features over score-independent ones [8]. Therefore, the design of deep architectures that incorporate score information is an obvious and overdue extension of previous approaches.

The goal of this paper is to explore different methods to incorporate this score information. Our hypothesis is that including score information will lead to improved performance of deep networks in the objective MPA task. To this end, we present three architectures which combine score and audio features to make a *score-informed* assessment of a music performance. First, we concatenate aligned pitch contours and scores into a 2-dimensional time-series feature representation that is fed to a convolutional neural network (CNN). Second, we propose a joint embedding model for aligned score and pitch contours. The assessment ratings are predicted using the cosine similarity between the score and performance embeddings. Third, we utilize the distance matrix, a mid-level representation combining both the score and pitch contour, as the input to a deep CNN trained to predict the teachers' assessments. Finally, using a fairly large scale dataset of middle school and high school student auditions, we perform an in-depth evaluation comparing these proposed architectures against each other and with a score-independent baseline approach for MPA .

2. RELATED WORK

MPA deals with the task of assessing music performances based on audio recordings. Progress in MPA is roughly categorized into feature design-based approaches [8, 13, 20, 24, 30] and feature learning-based approaches [9, 22, 31]. Feature design-based methods rely on signal processing techniques to either extract standard spectral and temporal features [13], or use expert knowledge to extract perceptually motivated features capable of characterizing music performances [20, 24]. The extracted features are then fed into simple machine learning classifiers to train models which predict different performance assessment ratings. Feature learning-based approaches, on the other hand, stem from the argument that important features for modeling performance assessments are not trivial and cannot be easily described. Hence, they rely on using mid-level representations (such as pitch contours or mel-spectrograms) as input to sophisticated machine learning models such as sparse coding [9, 31] and neural networks [22].

Most performances of Western music, however, are based on written musical scores. Hence, performances are also assessed based on their perceived deviations from the underlying score. There has been some prior research on incorporating the score information into the assessment modeling process. Most of the these approaches rely on computing descriptive features using some notion of *distance* between the score representation and the performance representation [3, 6, 11, 17, 19]. The most common approach has been to first use an alignment algorithm, e.g., Dynamic Time Warping (DTW) [25], to temporally align the performance recording with the score and then compute descriptive features which characterize the deviations of the performance from the score [1, 27]. However, to the best of our knowledge, incorporating score information directly into neural network-based models for MPA has not been investigated before.

Score-informed approaches have helped improve results for both related performance analysis tasks and other music information retrieval tasks. Most of these methods have also relied on an alignment between the audio recording and the score as the primary tool for incorporating score information. Aligning audio recordings with scores has been useful for several tasks such as detecting expressive features in music performances [15], identifying missing notes and errors in piano performances [5], and segmenting syllables in vocal performances [23]. Scores have also been used to generate soft labels and/or artificial training data for tasks such as source separation [4, 18].

3. METHODS

We propose and compare three different approaches to incorporate the score information with audio features for MPA. [1] The score information is represented as the MIDI pitch sequence (in ticks) obtained from the sheet music of the score to be performed. Henceforth, the MIDI pitch sequence will be referred to as the *score*. The student's *performance* is represented by the pitch contour of the audio. We use pitch contour since it captures both pitch and rhythmic information. Musical dynamics and timbre are ignored in this study; while dynamics are an important expressive tool for the performer [14], the score usually lacks specificity in dynamics instructions and cannot serve as the same absolute reference as for pitch and rhythm.

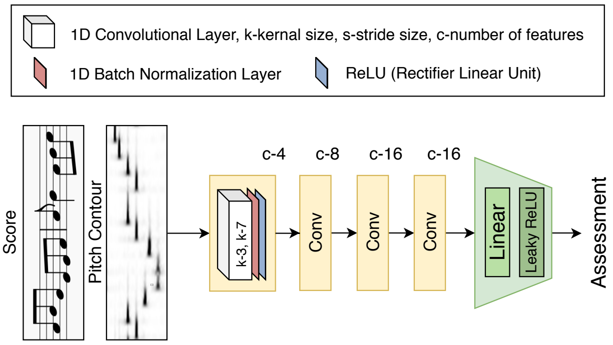

1D Convolutional Layer, k-kernal size, s-stride size, c-number of features
1D Batch Normalization Layer
ReLU (Rectifier Linear Unit)
Score
Pitch Contour
c-4
c-8
c-16
c-16
k-3, k-7
Conv
Conv
Conv
Linear
Leaky ReLU
Assessment

Figure 1. Schematic for the SIConvNet. The aligned *score* and *pitch contour* are stacked together and fed into a 4-layer CNN to directly predict the assessment ratings.

3.1 Score-Informed Network (SIConvNet)

The first approach that we use is probably the most straightforward way of incorporating score information into the assessment model. A simple CNN is used that relies on both the score and performance as the input and directly predicts the assessment ratings.

3.1.1 Input Representation

The input representation for this approach is constructed by simply stacking an aligned pitch contour and score pair to create a $N \times 2$ matrix, where N is the sequence length of the pitch contour. The two channels correspond to the pitch contour and score, respectively.

In order to obtain this representation, we first consider a pitch contour snippet of length N (sequence of logarithmic frequencies). Then, we find the corresponding part of the score using a DTW-based alignment process. The obtained score snippet (sequence of MIDI note numbers) is then resampled to have the same length N as the pitch contour.

3.1.2 Model Architecture

A schematic of the model architecture is shown in Figure 1. We use a simple 4-layer CNN based on the architecture proposed by Pati et al. [22] and append a single linear layer which predicts the assessment. Each convolutional stack consists of a 1-D convolution followed by a 1-D batch normalization layer [12] and ReLU non-linearity. The linear layer at the end comprises of a dense layer followed by Leaky ReLU non-linearity.

[1] The code is available at: https://github.com/biboamy/FBA-Fall19

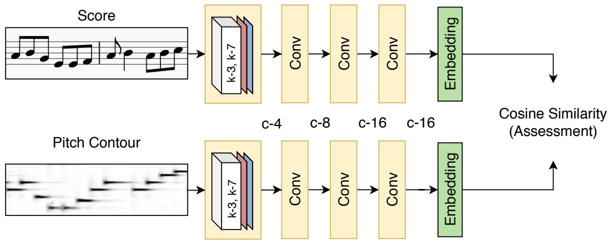


Figure 2. Schematic of the JointEmbedNet architecture.

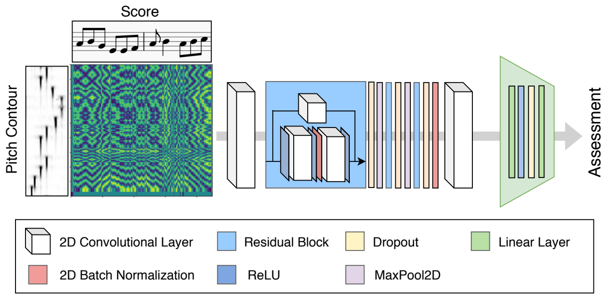


Figure 3. Schematic of the DistMatNet architecture.

3.2 Joint Embedding Network (JointEmbedNet)

The second approach is based on the assumption that performances are rated based on some sort of perceived distance between the performance and the underlying score being performed. Consequently, we use two separate encoder networks to project the score and the pitch contour to a joint latent space and then use the similarity between the embeddings to predict the assessment ratings.

3.2.1 Input Representation

This approach uses the same input representation as SIConvNet (see Section 3.1.1). However, instead of stacking the aligned pitch contour and the score, the individual $N \times 1$ sequences are fed separately to the two encoders.

3.2.2 Model Architecture

This network (see Figure 2) uses two 1-D convolutional encoders having the same architecture as SIConvNet. Each encoder has 4 convolutional blocks to extract high level feature embeddings. The performance encoder is expected to extract relevant features pertaining to the performance from the pitch contour. On the other hand, the score encoder is expected to extract the important features from the score. Assuming that the assessment rating for the performance is high if these two embeddings are similar, we use the cosine similarity $\cos(E_{\text{score}}, E_{\text{performance}})$ between the two embeddings to obtain the predicted assessment rating. E_{score} and $E_{\text{performance}}$ are the embeddings obtained from the score and performance encoders, respectively. If the two embeddings are similar, the cosine similarity is close to one, and the model will predict a higher rating.

3.3 Distance Matrix Network (DistMatNet)

The final approach uses a distance matrix between the pitch contour and the score as the input to the network. Given the information from both the pitch contour and the score, the task of performance assessment might be interpreted as finding a *performance distance* between them. Thus, the choice of the distance matrix as the input representation allows the model to learn from the pitch differences. A Residual CNN [10] architecture is chosen for the network.

3.3.1 Input Representation

The distance matrix elements are the pair-wise wrapped distances between the pitch contour and the MIDI pitch sequence. The octave-independent wrapped-distance is used to compensate for possible octave errors made by the pitch tracker. To ensure a uniform input size to the network, the matrix is resampled to a square shape of a fixed size. Thus, a performance with constant tempo would result in an aligned path located on the diagonal. Unlike the previous two methods where the input pitch contour and the score are aligned using DTW, the distance matrix input avoids any error propagation caused by alignment errors. The choice of this input representation stems from the success of distance matrices (or self-similarity matrices) in other areas of MIR such as structural segmentation [2, 7] and music generation [28].

3.3.2 Model Architecture

The model architecture is shown in Figure 3. It is composed of 3 residual blocks. Each residual block has 2 convolutional layers. Dropout and max-pooling are added between each residual block. A classifier with two linear layers (128 features) with one ReLU and dropout layer in between is used after the residual network to perform regression prediction. We use (3,3) kernal size and 4 feature maps for all convolutional layers, 0.2 dropout rate, and a (3,3) kernal size for all max-pooling layers.

4. EXPERIMENTS

4.1 Dataset

The dataset we use to evaluate our methods is a subset of a large dataset of middle school and high school student performances. These are recorded for the Florida All State Auditions, which are separated into three bands: (i) middle school band, (ii) concert band, and (iii) symphonic band. The recordings contain auditions spanning 6 years (from 2013 to 2018), and feature several monophonic pitched and percussion instruments. Each student performs rehearsed scores, scales and a sight reading exercise. For the purpose of this study we limit our experiments to the *technical etude* for middle school and symphonic band auditions. We choose *Alto Saxophone*, *Bb Clarinet* and *Flute* performances due to these being the most popular across all pitched instruments. Table 1 shows the distribution of data across different instruments. The average duration of each performance is 30 s for middle school and 50 s for symphonic band students. The dataset also includes the musical scores that the students are supposed to perform for each exercise. The average length (in notes) of the musical

	Middle School	Symphonic Band
Alto Saxophone	696	641
Bb Clarinet	925	1156
Flute	989	1196

Table 1. Number of performances for the different instruments per band.

scores are 136 for middle school and 292 for symphonic band. Note that these scores are the same across all students performing the same instrument in the same year but vary across years and instruments.

The dataset also contains expert assessments for each exercise of a student performance. Each performance is rated by one expert along 4 criteria defined by the Florida Bandmasters' Association (FBA): (i) musicality, (ii) note accuracy, (iii) rhythmic accuracy, and (iv) tone quality. All ratings are on a point-based scale and are normalized to range between 0 to 1 by dividing by the maximum. Since we focus on pitch contours as the primary audio feature, tone quality is excluded from this study.

4.1.1 Data pre-processing

The pitch contours are extracted using the pYIN algorithm [16] with a block size and hop size of 1024 and 256 samples, respectively. The audio sampling rate is 44100 Hz. The extracted frequencies are converted from Hz to MIDI pitch (unlike the MIDI pitches from the musical score, these can be floating point numbers). Both the resulting pitch contour and musical score are normalized by dividing by 127. Finally, for the purpose of model training and evaluation, we divide our dataset into three randomly sampled subsets: training, validation, and testing. We use a ratio of $8\colon 1\colon 1$ for splitting the dataset.

We use *random-chunking* as a data augmentation tool when training SIConvNet and JointEmbedNet since it has shown to be useful in improving model performance [22]. First, the pitch contour is chunked into snippets of length N by randomly selecting the starting position. The corresponding aligned and length-adjusted score snippet is obtained using the method described in Section 3.1.1. We assume the chunked segment has the same assessment score as the whole recording. We do not perform chunking on our distance matrix since the matrix has already been resampled into a smaller resolution. Instead, we discuss how varying the resampling size could effect the performance in one of the experiments.

4.2 Experimental Setup

We present three experiments to evaluate our proposed methods. First, we compare the overall performance of the proposed architectures against a score-independent baseline system PCConvNet [22] which uses only the randomly-chunked pitch contour as input. This experiment also gives us an indication of the effectiveness of each of the proposed methods. Second, we look at the sensitivity of the SIConvNet and JointEmbedNet to the chunk size N. Finally, we investigate the effect of varying the resolution of the input distance matrix for the DistMatNet model. The latter two experiments were aimed at understanding the effects of the different hyper-parameters used while constructing the input data for each model. These helped us arrive at the best parameters for each approach.

SIConvNet	JointEmbedNet	DistMatNet
3,089	6,144	63,417

Table 2. Number of parameters for each model.

The number of trainable parameters for each method is shown in Table 2. DistMatNet has a higher number of parameters because it uses a higher-dimensional input with a deeper architecture to capture high level information [10].

For each method, we trained separate models to predict each assessment criterion. Moreover, to measure the variation of each model, we trained each model on 10 different random seeds. We used M_i to represent the model training on different random seed where $i = 0 \ldots 9$. A boxplot with median and variation of each M_i is shown to demonstrate the results. All the models are trained based on the mean squared error between estimated and ground truth ratings. All the models are trained with a stochastic gradient descent optimizer with a 0.05 learning rate. We apply early stopping if the validation loss does not improve for 100 epochs. The performance of all models is measured using the coefficient of determination (R^2 score):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \bar{y_i})^2}, \qquad (1)$$

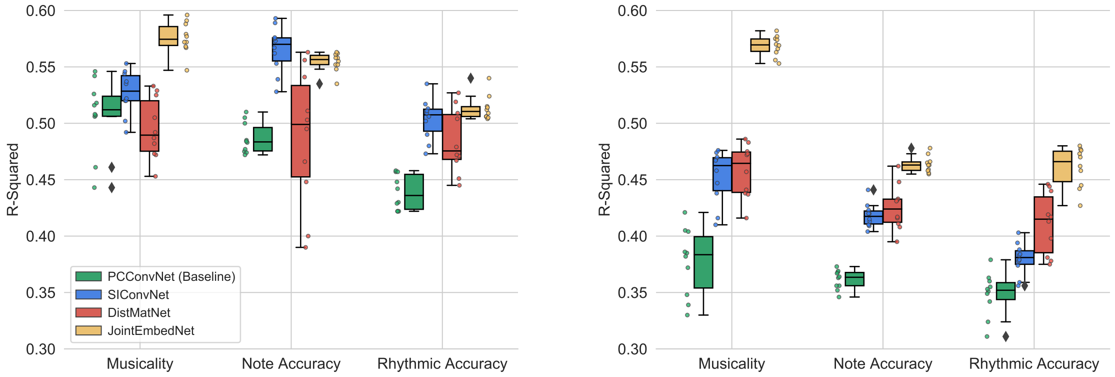
where y_i is the ground truth rating, $\hat{y_i}$ is the estimated rating, and $\bar{y_i}$ is the average of the ground truth rating. R^2 is a common metric to evaluate the fit of a regression prediction to the ground truth value.

5. RESULTS & DISCUSSION

5.1 Overall Performance

Figure 4 shows the comparative performance for all models for middle school and symphonic band. We can make the following observations (with independent t-test results reported):

(i) We compare the performance of various models trained on different band performances. All systems perform better (higher R^2 value) on the middle school recordings than on the symphonic band recordings ($p < 0.01$ except JointEmbedNet for musicality). One possible explanation for this is that symphonic band scores are usually more complicated and longer. For example, symphonic band scores tend to be performed at high tempo with high note density. The chunking into smaller lengths (and the downsampling of the distance matrix) compared to the score length might lead to a less accurate mapping to the assessment rating. An additional factor is that most performers in the symphonic band auditions exhibit greater skill level than middle school performers thus making it



(a) Middle School

(b) Symphonic Band

Figure 4. Box plots showing comparative performance (higher is better) across different models and assessment criteria.

potentially more difficult to model the differences in proficiency levels.

(ii) All score-informed models generally outperform the baseline, implying that score information is indeed helpful for MPA ($p < 0.01$ except SIConvNet for musicality, DistMatNet for musicality and note accuracy on middle school). We notice, however, that the difference between the score-independent baseline and the score-informed models is smaller for the middle school than for symphonic band. Given the significant improvement over the baseline for symphonic band performances (which have complicated scores), we speculate that the score-informed models benefit more from access to score information. In other words, a score reference becomes more impactful with increasing proficiency level while the pitch contour alone contains most relevant information for medium proficiency levels.

(iii) While the two models SIConvNet and JointEmbedNet both use the same input features, JointEmbedNet either outperforms or matches SIConvNet in all experiments. The main difference between these two architectures is that SIConvNet simply performs a regression to estimate the assessments while JointEmbedNet learns a similarity in the embedding space to model the assessments. Therefore, we can assume that JointEmbedNet is able to explicitly model the differences between the input pitch contour and score especially in the case of symphonic band where the scores are more complicated.

(iv) We observe that while DistMatNet and JointEmbedNet both utilize the similarity between the score and pitch contour, albeit at different stages of the network, JointEmbedNet typically performs better across categories and bands, and the gap is larger for musicality than for the other two categories. It is possible that the absolute pitch at the input may be important for the final assessment (octave jumps, for example, would not be properly modeled in the distance matrix). More likely, however, is that the significantly larger input dimensionality of the matrix (compared to the aligned sequences for JointEmbedNet) negatively impacts performance. Most of the relevant information for MPA centers around the diagonal of the distance matrix with relatively small deviations depending on the students' tempo variation. Most of the distance matrix elements far from the diagonal contain redundant or irrelevant information, thus complicating the task. Another advantage that JointEmbedNet might have over DistMatNet in terms of overall assessment is that the distance is computed on the whole performance while DistMatNet computes a frame-level pitch distance, potentially complicating the task for overall quality measures like musicality.

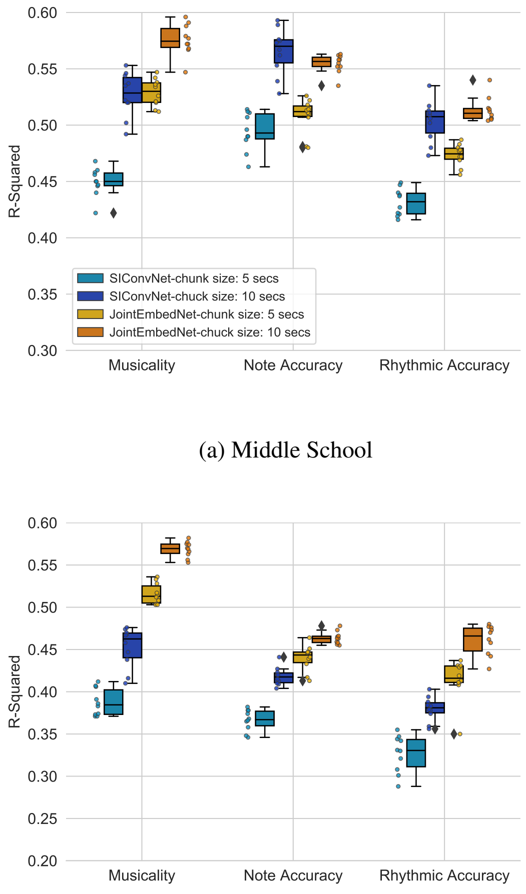
5.2 Chunk Size

In this experiment, we look at the impact of two different chunk sizes for the first two methods. Figure 5 shows the results on middle school (a) and symphonic band (b). For both SIConvNet and JointEmbedNet, a chunk size of 10 s outperforms that of 5 s across all the bands.

Chunking with random sampling is a form of data augmentation. By using the ground truth rating of the whole performance, the chunks are assumed to reflect the quality of the whole performance. The results show that 5 s chunks might be too short to evaluate the whole performance while 10 s chunks are much better suited regardless of category and score complexity. Chunk lengths greater than 10 s were not tested because we restricted ourselves to the length of the shortest performance in the dataset. Consequently, we used a 10 s chunk size for the experiment in Figure 4.

5.3 Distance Matrix Resolution

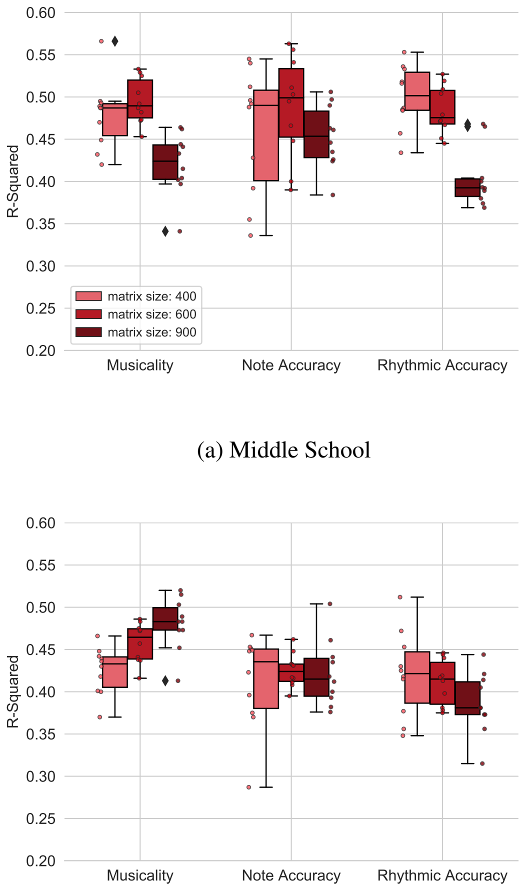
In this experiment, we study the impact of the different input matrix resolutions 400×400, 600×600, and 900×900, for the DistMatNet model. The results for both middle school and symphonic band are shown in Figure 6. First, the performance of rhythmic accuracy criterion tends to decrease with increasing distance matrix resolution. It might



(a) Middle School

(b) Symphonic Band

Figure 5. Box plots showing comparative performance (higher is better) across different chunk sizes for SIConvNet and JointEmbedNet.



(a) Middle School

(b) Symphonic Band

Figure 6. Box plots showing comparative performance (higher is better) across different matrix sizes for the distance matrix network (DistMatNet).

be more difficult for the same model structure to capture the complexity inside a larger matrix. This can also explain the result for middle school: although increasing the input resolution from 400 × 400 to 600 × 600 will capture more details, the performance decreases when the matrix resolution is further increased. Second, an input matrix size of 600 × 600 leads to a slightly higher average score (0.46) on both symphonic and middle school than the other two resolutions (0.45 for 400 × 400 and 0.43 for 900 × 900). We ended up using the 600 × 600 resolution for the experiment in Figure 4.

6. CONCLUSION

This paper presents three novel neural network-based methods that combine score information with a pitch representation of an audio recording to assess a music performance. The methods include: (i) a CNN with aligned pitch contour and score as the input, (ii) a joint embedding model that learns the assessment as the cosine similarity of the embeddings of both the aligned pitch contour and the score, and (iii) a distance-matrix based CNN, using a differential representation of pitch contour and score at the input. The results show that all the methods outperform the score-independent baseline model. The joint embedding model achieves the highest average performance.

Beyond the obvious applications in software-based music tutoring systems, score-informed performance assessment models (and objective MPA in general) can benefit the broader area of music performance analysis. Models capable of rating performances along different criteria could serve as useful tools for objective evaluation of generative systems of music performance. In addition, such models could also be explored for objective analysis of inter-annotator differences in rating music performances.

In the future, we plan to incorporate timbre and dynamics information into the models as it has been shown to improve accuracy [22]. This will also enable the model to assess performances in terms of tone quality, the criterion ignored in this study. We also plan to investigate other instruments and to examine cross-instrument relationships by training instrument-specific models. Furthermore, the musical score reference could be replaced with other representations such as the pitch contour of a highly-rated performance.

7. ACKNOWLEDGMENTS

Jiawen Huang and Yun-Ning Hung are the main contributors and contributed equally to this work.

We would like to thank the Florida Bandmasters Association for providing the dataset used in this study. We also gratefully acknowledge Microsoft Azure who supported this research by providing computing resources via the Microsoft Azure Sponsorship.

8. REFERENCES

[1] B. Bozkurt, O. Baysal, and D. Yuret. A dataset and baseline system for singing voice assessment. In *Proc. of International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 430–438, Matosinhos, Porto, 2017.

[2] Alice Cohen-Hadria and Geoffroy Peeters. Music structure boundaries estimation using multiple self-similarity matrices as input depth of convolutional neural networks. In *Proc. of Audio Engineeing Society (AES) International Conference on Semantic Audio*, Erlangen, Germany, 2017.

[3] Johanna Devaney, Michael I Mandel, and Ichiro Fujinaga. A Study of Intonation in Three-Part Singing using the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT). In *Proc. of 13th International Society of Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, 2012.

[4] Sebastian Ewert and Mark B Sandler. Structured dropout for weak label and multi-instance learning and its application to score-informed source separation. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2277–2281, New Orleans, USA, 2017.

[5] Sebastian Ewert, Siying Wang, Meinard Müller, and Mark Sandler. Score-informed identification of missing and extra notes in piano recordings. In *Proc. of 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York City, USA, 2016.

[6] Felipe Falcao, Baris Bozkurt, Xavier Serra, Nazareno Andrade, and Ozan Baysal. A Dataset of Rhythmic Pattern Reproductions and Baseline Automatic Assessment System. In *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[7] Thomas Grill and Jan Schluter. Music boundary detection using neural networks on spectrograms and self-similarity lag matrices. In *Proc. of 23rd European Signal Processing Conference (EUSIPCO)*, pages 1296–1300, Nice, France, 2015.

[8] Siddharth Gururani, Ashis Pati, and Alexander Lerch. Analysis of objective descriptors for music performance assessment. In *Proc. of International Conference on Music Perception and Cognition (ICMPC)*, Montréal, Canada, 2018.

[9] Yoonchang Han and Kyogu Lee. Hierarchical approach to detect common mistakes of beginner flute players. In *Proc. of 15th International Society of Music Information Retrieval Conference (ISMIR)*, pages 77–82, Taipei, Taiwan, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, USA, 2016.

[11] Jiawen Huang and Alexander Lerch. Automatic assessment of sight-reading exercises. In *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. of 32nd International Conference on Machine pyin (ICML)*, pages 448–456, Lille, France, 2015.

[13] Trevor Knight, Finn Upham, and Ichiro Fujinaga. The potential for automatic assessment of trumpet tone quality. In *Proc. of 12th International Society of Music Information Retrieval Conference (ISMIR)*, pages 573–578, Miami, USA, 2011.

[14] Alexander Lerch, Claire Arthur, Ashis Pati, and Siddharth Gururani. Music performance analysis: A survey. In *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[15] Pei-Ching Li, Li Su, Yi-Hsuan Yang, Alvin WY Su, et al. Analysis of expressive musical terms in violin using score-informed and expression-based audio features. In *Proc. of 16th International Society of Music Information Retrieval Conference (ISMIR)*, pages 809–815, Málaga, Spain, 2015.

[16] Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, Florence, Italy, 2014.

[17] Oscar Mayor, Jordi Bonada, and Alex Loscos. Performance analysis and scoring of the singing voice. In *Proc. of Audio Engineering Society (AES) Convention*, pages 1–7, London, UK, 2009.

[18] Marius Miron, Jordi Janer Mestres, and Emilia Gómez Gutiérrez. Monaural score-informed source separation for classical music using convolutional neural networks. In *Proc. of 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017.

[19] Emilio Molina, Isabel Barbancho, Emilia Gómez, Ana Maria Barbancho, and Lorenzo J Tardón. Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 744–748, Vancouver, Canada, 2013.

[20] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Proc. of International Conference on Spoken Language Processing (INTERSPEECH)*, pages 1706–1709, Pittsburg, USA, 2006.

[21] Caroline Palmer. Music performance. *Annual review of psychology*, 48(1):115–138, 1997.

[22] Ashis Pati, Siddharth Gururani, and Alexander Lerch. Assessment of student music performances using deep neural networks. *Applied Sciences*, 8(4):507, 2018.

[23] Jordi Pons Puig, Rong Gong, and Xavier Serra. Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks. In *Proc. of 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017.

[24] Oriol Romani Picas, Hector Parra Rodriguez, Dara Dabiri, Hiroshi Tokuda, Wataru Hariya, Koji Oishi, and Xavier Serra. A real-time system for measuring sound goodness in instrumental sounds. In *Proc. of Audio Engineering Society Convention*, Philadelphia, USA, 2015.

[25] Hiroaki Sakoe and Seibi Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[26] Sam Thompson and Aaron Williamon. Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception: An Interdisciplinary Journal*, 21(1):21–41, 2003.

[27] Amruta Vidwans, Siddharth Gururani, Chih-Wei Wu, Vinod Subramanian, Rupak Vignesh Swaminathan, and Alexander Lerch. Objective descriptors for the assessment of student music performances. In *Proc. of Audio Engineeing Society (AES) International Conference on Semantic Audio*, Erlangen, Germany, 2017.

[28] I-Chieh Wei, Chih-Wei Wu, and Li Su. Generating strcutured drum pattern using variational autoencoder and self-similarity matrix. In *Proc. of 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019.

[29] Brian C Wesolowski, Stefanie A Wind, and George Engelhard. Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted rasch partial credit model. *Music Perception: An Interdisciplinary Journal*, 33(5):662–678, 2016.

[30] Chih-Wei Wu, Siddharth Gururani, Christopher Laguna, Ashis Pati, Amruta Vidwans, and Alexander Lerch. Towards the objective assessment of music performances. In *Proc. of International Conference on Music Perception and Cognition (ICMPC)*, pages 99–103, San Francisco, USA, 2016.

[31] Chih-Wei Wu and Alexander Lerch. Learned features for the assessment of percussive music performances. In *Proc. of International Conference on Semantic Computing (ICSC)*, Laguna Hills, California, USA, 2018.