Score-Informed Networks for Music Performance Assessment

Jiawen Huang*
Queen Mary University of London
London, UK

Yun-Ning Hung*, Ashis Pati, Siddharth Gururani, Alexander Lerch
Georgia Institute of Technology
Atlanta, GA, USA

Georgia Tech Center for Music Technology

INTRODUCTION

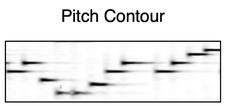
- **Incorporate score information into deep neural network-based methods for music performance assessment**
- **Propose three network architectures**
  - **Score-Informed Network:** a CNN that utilizes a 2-dimensional time-series input comprising of aligned pitch contours and score
  - **Joint Embedding Network:** a joint embedding model which learns a joint latent space for pitch contours and scores
  - **Distance Matrix Network:** a distance matrix-based residual CNN which utilizes patterns in the distance matrix between pitch contours and musical score to predict assessment ratings
- **Compare to a score-independent baseline [22]**

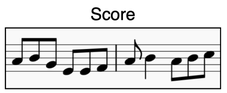
PREPROCESSING

- **Pitch Contour**
  - Extracted using pYin [16]
  - Converted from Hz to MIDI pitch
- **Normalization**

$$pc_{norm,i} = \frac{pc_i}{127}$$

$$sc_{norm,i} = \frac{sc_i}{127}$$





DATASET

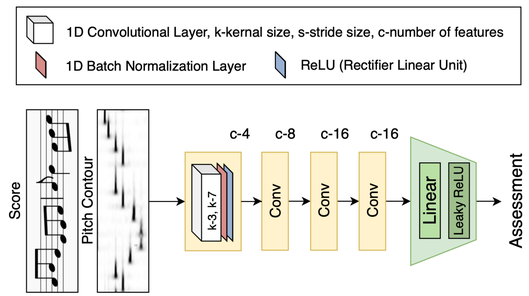
- **Recordings at auditions with expert assessment**
  - 2 bands of students: middle school band and symphonic band
  - 3 instruments: Alto Saxophone, Bb Clarinet and Flute
  - 5603 performances, 18 different pieces
- **Three assessment categories**
  - Musicality
  - note accuracy
  - rhythmic accuracy
- **Provided by Florida Bandmasters Association (FBA)**

* These authors contributed equally to this work.

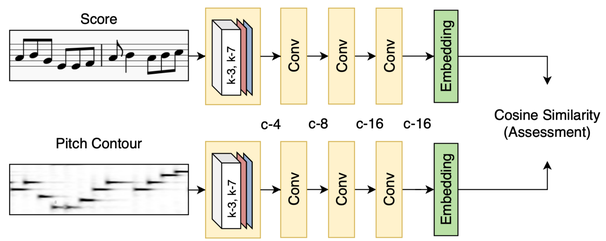
[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[16] Mauch, M., & Dixon, S. (2014, May). pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 659-663). IEEE.
[22] Pati, K. A., Gururani, S., & Lerch, A. (2018). Assessment of student music performances using deep neural networks. Applied Sciences, 8(4), 507.

NETWORK ARCHITECTURES

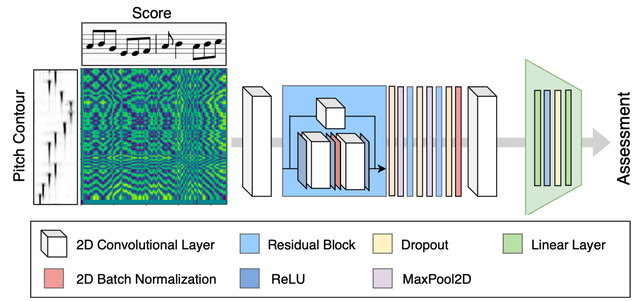
- **Score-Informed Network (SIConvNet)**
  - Score and pitch contour snippets stacked together
  - 4-layer CNN to directly predict the assessment



- **Joint Embedding Network (JointEmbedNet)**
  - Score and pitch contour snippets projected to a joint latent space
  - Similarity between the embeddings as the assessment

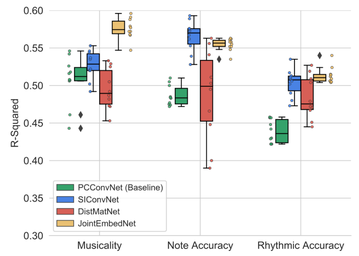


- **Distance Matrix Network (DistMatNet)**
  - Distance matrix between the score and the pitch contour as the input
  - A Residual CNN [10] to find the performance distance

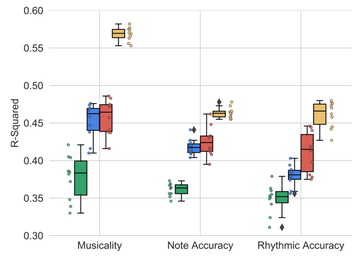


RESULTS AND DISCUSSION

- **Score-informed vs. Score-independent baseline**
  - Score-informed models generally outperform the baseline
- **Middle school vs. Symphonic band**
  - All systems perform better on middle school recordings
- **JointEmbedNet vs. SIConvNet**
  - Use the same input features
  - JointEmbedNet outperforms or matches SIConvNet
- **JointEmbedNet vs. DistMatNet**
  - Both utilize the similarity between score and pitch contour
  - JointEmbedNet peforms better across categories and bands

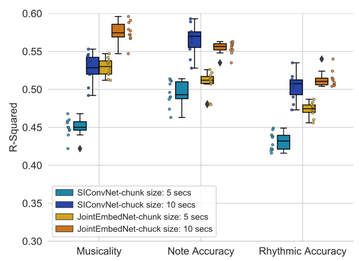


(a) Middle School

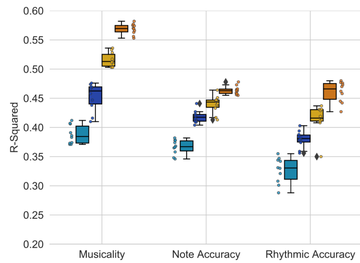


(b) Symphonic Band

- **Random-Chunking for SIConvNet and JointEmbedNet**
  - Assumption: chunks reflect the quality of the whole performance
  - 10 sec chunks are better suited than 5 sec regardless of category and score complexity



(a) Middle School



(b) Symphonic Band

CONTACT

Jiawen Huang
Center for Digital Music
Queen Mary University of London
jiawen.huang@qmul.ac.uk

Github Repository