# SUPPLEMENTARY DOCUMENT FOR GRAPH NEURAL NETWORK FOR MUSIC SCORE DATA AND MODELING EXPRESSIVE PIANO PERFORMANCE

**Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Juhan Nam**

May 2019

## 1 Feature Detail

The list of input features are: **a)** pitch in MIDI pitch(int) **b)** duration **c)** pitch in octave (int) and pitch class as one-hot 12-D vector **d)** beat importance in measure (rule-based) **e)** measure length **f)** duration of following rest **g)** duration after the corresponding tempo marking **h)** duration after the corresponding dynamic marking **i)** relative position in measure [0 - 1] **j)** relative position in entire piece [0 - 1] **k)** if grace note, distant to its following non-grace note (int) **l)** is preceded by a grace note (bool) **m)** time signature denominator vector (4-D vector: 2, 4, 8, 16) **n)** time signature numerator vector (5-D vector: duple, triple, quadruple, compound, other) **o)** tempo marking vector (5-D vector) **p)** dynamic marking vector (4-D vector) **q)** slur and beam status vector (6-D multi-hot vector: start, end, continue for slur and beam) **r)** composer vector (16-D one-hot vector) **s)** notation marking vector for trill, fermata, accented, strong accented, staccato, tenuto, arpeggiate, cue (8-D multi-hot vector)

Every duration or length is measured in a quarter note.

These global conditioning features are added to input features along with the corresponding score features: **a)** embedded tempo marking vector of the beginning of the piece **b)** tempo of the first ten beat in quarter note per minute (log)

The global conditions are also directly concatenated with the input of the LSTM that predicts tempo in the performance decoder.

The output features are: **a)** Tempo in quarter note per minute, calculated for every beat (log) **b)** MIDI velocity **c)** deviation of the note onset compared to 'in-tempo position' in quarter note **d)** articulation (log) **e)** value of sustain pedal at note onset [0 - 127] **f)** value of sustain pedal at note offset [0 - 127] **g)** smallest sustain pedal value between the note's onset and the offset [0 - 127] **h)** elapsed time between the note onset and the moment g) is detected **i)** smallest sustain pedal value between the note offset and its closest next note onset **j)** elapsed time between the note offset and the moment i) is detected **k)** value of soft pedal at note onset.

We used additional features for modeling a trill note. The trill features are **a)** number of trill notes per second **b)** relative velocity of the last trill note compared to the first trill note **c)** relative duration ratio of the first trill note **d)** relative duration ratio of the last trill note **e)** whether the trill starts with higher note (alternative trill).

We trained an additional neural network to predict trill features for the notes with trill mark.

## 2 Model Parameters

1. Baseline Model: 3L 256D Bi-LSTM for score encoder and 1L 128D auto-regressive Uni-LSTM for decoder. Trained with maximum KLD weight of 0.02.
2. HAN Model: 2L 128D Bi-LSTM for note-wise, 2L 128D Bi-LSTM for voice-wise, 2L 128D Bi-LSTM for beat-level, 1L 128D Bi-LSTM for measure-level, and 1L 128D auto-regressive Uni-LSTM for decoder. Trained with maximum KLD weight of 0.0003.

3. G-HAN Model: 2L 192D 12Edge GGNN for note-level, 2L 128D Bi-LSTM for beat-level, 1L 64D Bi-LSTM for measure-level, and 1L 128D auto-regressive Uni-LSTM for decoder. Trained with maximum KLD weight of 0.0003.
4. Proposed ISGN Model: 2L 128D 12Edge GGNN for note-level, 1L 32D Bi-LSTM for measure-level in score encoder. 1L 12E 32Margin GGNN with 1L 32D beat-level Bi-LSTM for decoder. Trained with maximum KLD weight of 0.003.

The codes and pre-trained models are available in https://github.com/jdasam/virtuosoNet

# 3 Details in Training

## 3.1 Data Cleaning

Our dataset consists of automatically-aligned pairs of XML score and human performance MIDI. The automatic alignment included some errors, which matched the score notes to the performance notes in a different beat or measure. Among other output features, the onset deviation is hugely vulnerable to this type of error. Therefore, we cleaned the alignment result by omitting notes showing too large onset deviation.

In our first submitted version of the paper, we did not clean the alignment errors. In our camera-ready version, we have cleaned these errors and the cleaning made a significant improvement in the result. Without the cleaning, the onset deviation of the Baseline and the HAN model became too large. On the other hand, G-HAN and ISGN were robust to alignment error and ISGN made a perceptually better result compared to the Baseline and HAN model in our previous listening test. The gap between ISGN and HAN or Baseline in the new listening test was reduced after the data cleaning.

## 3.2 Weight for Features

Since our system predicts the different type of features at once, the result of training can be largely varied by how we define the loss weight for each feature. For example, if we give more weight in the tempo feature, the trained system can predict the tempo accurately, but the predictions for other features like velocity or pedal can be less accurate. We have tried several weight combinations and decided to give equal weight for each output feature. Finding the optimal weight combination or designing the system as a multitask have remained for future work.

Among the features, we applied different weight by note for the articulation (note duration) loss. The articulation of each note is largely affected by pedal usage. For example, if the sustain pedal is pressed at the note offset, the actual length of note sound is decided by the following pedal offset, not by the note offset itself. Therefore, we reduced the weight of articulation loss for notes with the sustain pedal pressed in the note offset to 0.1. By doing so, the articulation loss focuses more on the notes without sustain pedal at notes offset. Without this weight compensation, the overall note articulations become too short even for the notes without pedal.

# 4 Listening test details

## 4.1 Test pieces

The list of test pieces for experiments are :

1. Schubert, Franz - Piano Sonata No. 13, D.664 1st movement, beginning section
2. Chopin, Frédéric - Barcarolle, mm. 6-15
3. Liszt, Franz - Transcendental Études No. 5, mm. 84-102
4. Bach, Johann Sebastian - Prelude in F-sharp major, BWV 858, beginning section
5. Haydn, Joseph - Keyboard Sonata in E major, Hob.XVI:31, 1st movement, beginning section
6. Beethoven, Ludwig van -Piano Sonata No. 27, Op. 90, 1st movement, beginning section

Each pieces are cut into 30 seconds length for the experiments including theme of piece. Exact files used for experiments are in supplementary files.

| Schubert | | | | Chopin | | | |
|---|---|---|---|---|---|---|---|
| Human | ISGN | Baseline | HAN | Human | ISGN | Baseline | HAN |
| 9 | **11** | | | **17** | 3 | | |
| **17** | | 3 | | **19** | | 1 | |
| **17** | | | 3 | **15** | | | 5 |
| | 9 | **11** | | | 9 | **11** | |
| | 8 | | **12** | | 10 | | 10 |
| | | **11** | 9 | | | 4 | **16** |
| Liszt | | | | Bach | | | |
| Human | ISGN | Baseline | Han | Human | ISGN | Baseline | HAN |
| 4 | **16** | | | **19** | 1 | | |
| 5 | | 15 | | **18** | | 2 | |
| 8 | | | **12** | **14** | | | 6 |
| | **11** | 9 | | | 10 | 10 | |
| | **14** | | 6 | | 8 | | **12** |
| | | **12** | 8 | | | **12** | 8 |
| Haydn | | | | Beethoven | | | |
| Human | ISGN | Baseline | HAN | Human | ISGN | Baseline | HAN |
| 9 | **11** | | | 10 | 10 | | |
| **14** | | 6 | | **12** | | 8 | |
| **11** | | | 9 | **16** | | | 4 |
| | **12** | 8 | | | 6 | **14** | |
| | **12** | | 8 | | **13** | | 7 |
| | | 10 | 10 | | | **16** | 4 |

Table 1: Number of wins in pairwise listening test.

### 4.2 Piecewise results

Number of wins for all pairs are shown in Table1. Especially, the results between humans and models are also illustrated in the Figure1. Each pair is evaluated by 20 participants. In Chopin and Bach, human performance is dominant, but ISGN performances show similar or higher preferences in other works.

## 5 Tempo curve correlation

In Figure2-5, some of the examples of tempo curve of performances by human or generated by our models are presented. Each tempo value of performance is normalized with average tempo of the performance in the excerpts.

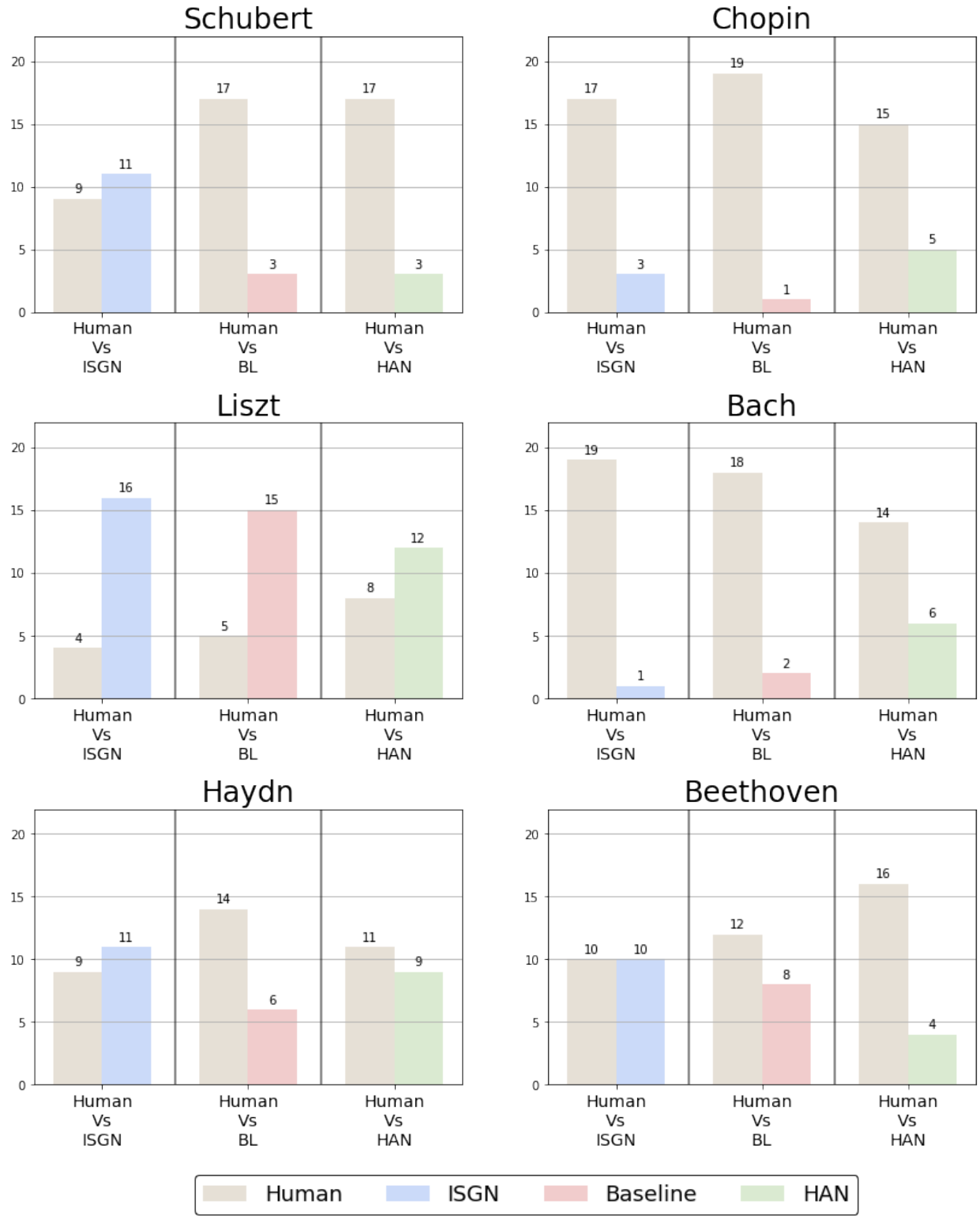


Figure 1: Piecewise human listening test results. The number on the bar graph represents the number of times a person or model has won.

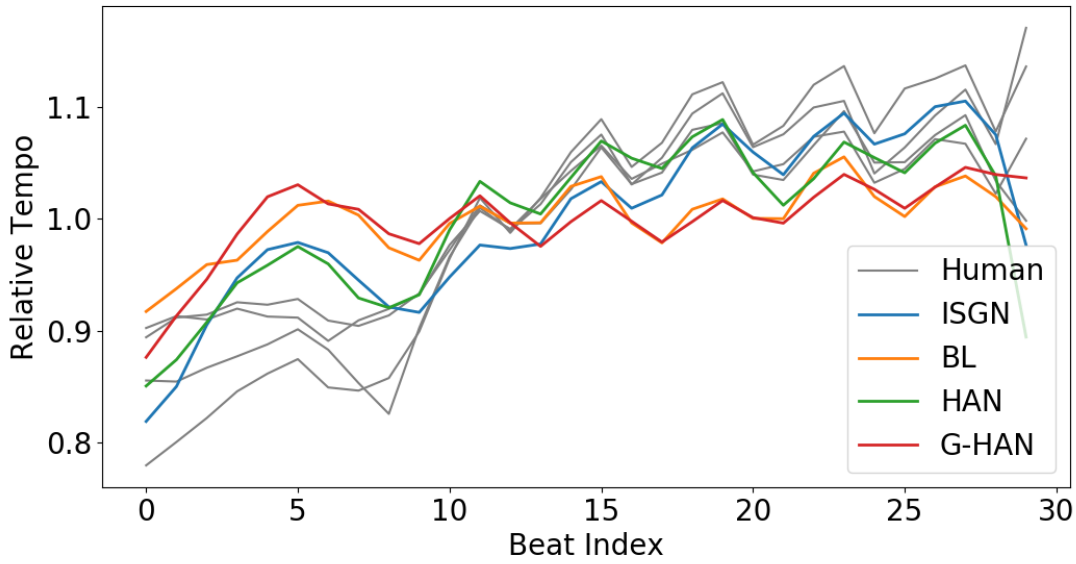


Figure 2: Tempo curves of performances of Schubert sonata D. 664, 1st movement, mm. 55-62

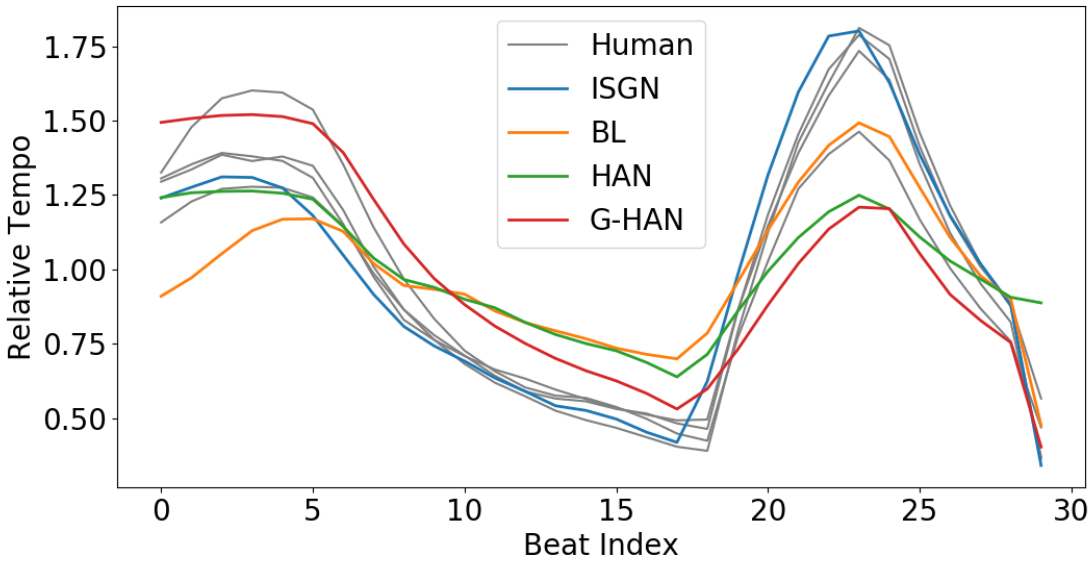


Figure 3: Tempo curves of performances of Beethoven Piano Sonata No. 17, 1st movement, mm. 226-241

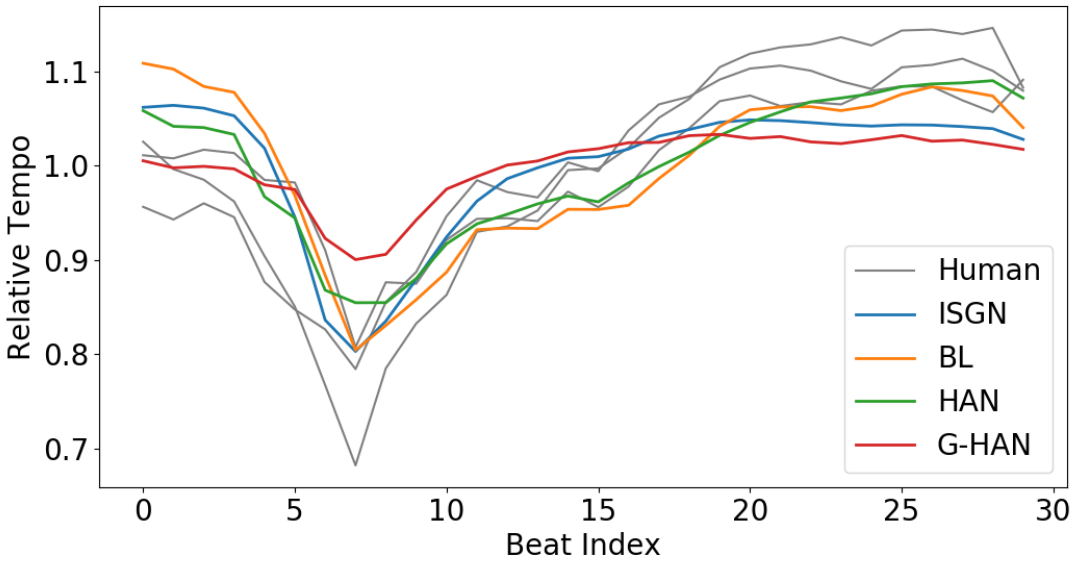


Figure 4: Tempo curves of performances from Haydn Keyboard Sonata No. 49, 1st movement, mm.10-15

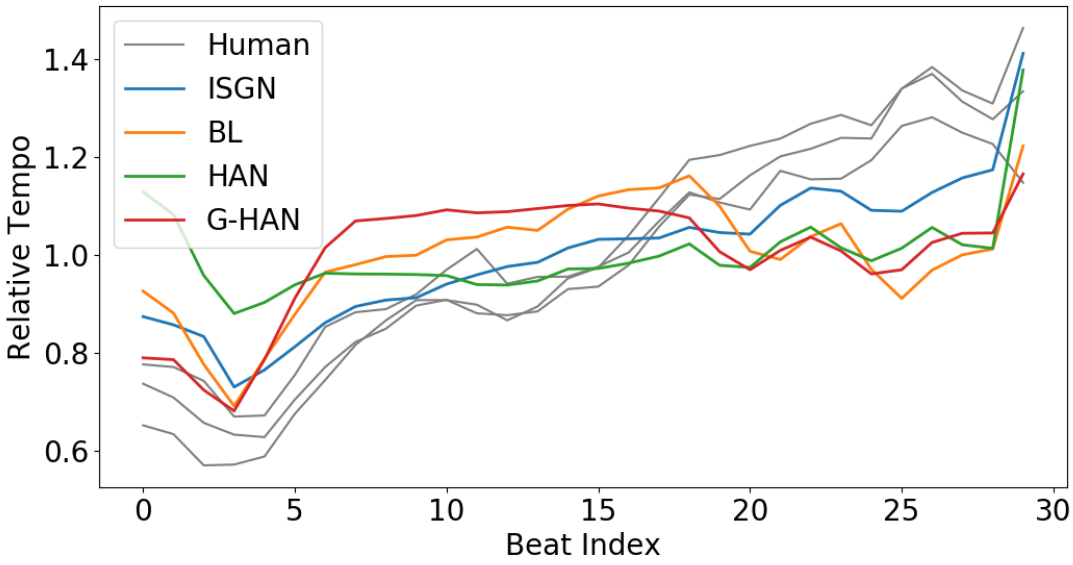


Figure 5: Tempo curves of performances of Liszt Transcendental Etude No. 9, mm. 53-66