The Jazz Transformer on the Front Line:
Exploring the Shortcomings of AI-Composed Music through Quantitative Measures

Shih-Lun Wu[1, 2] Yi-Hsuan Yang[2, 3]
<b06902080@csie.ntu.edu.tw> <yang@citi.sinica.edu.tw>
[1] National Taiwan University [2] Taiwan AI Labs [3] Academia Sinica. **Taipei, Taiwan.**

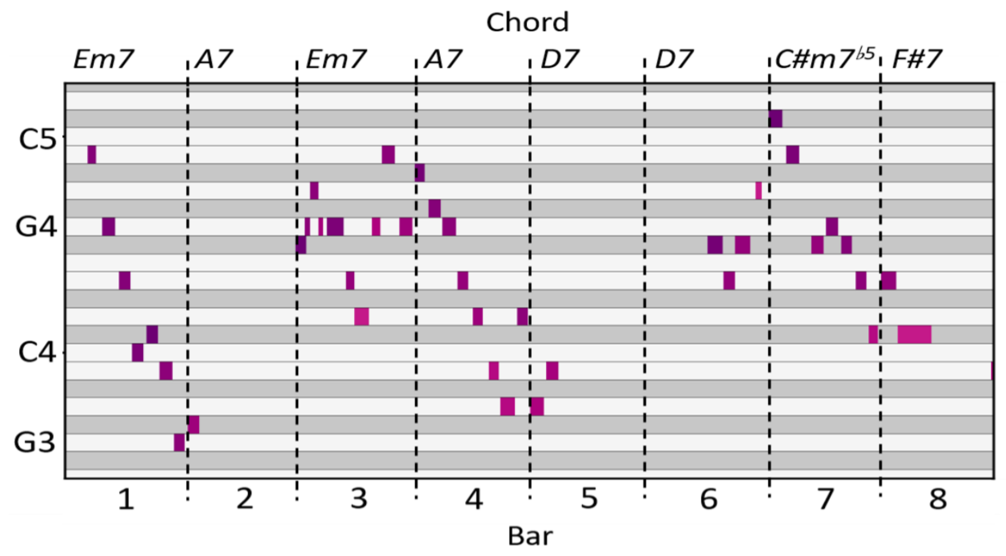

Fig. 1. *The first 8 bars of a piece composed by our **Model (B)***, in which we may see:

- **Clear rests between phrases**
- **Good combination of chords and melody.**

1. Motivation

The recent success of *Transformer* models in NLP and ample symbolic music datasets led to a new wave of research in **automatic music composition**. This boom, however, poses to us some great questions:

- The competence of *Transformers* is often claimed in the literature. **Is that the true story?**
- If not, can we **find the culprits in a quantitative manner?** (*i.e.*, not fully relying on user studies.)
- Do the **structure-related labels** (*e.g.*, phrases, parts) in *WJazzD* dataset **assist in models' learning?**

2. Building the Jazz Transformer

To feed Jazz music into the *Transformer* for training, first, it must be converted to a series of event tokens. Here is how we construct the vocabulary:

Note

NOTE-VELOCITY (32), NOTE-ON (128), NOTE-DURATION (32)

Metric

BAR, POSITION (64), TEMPO-CLASS (5), TEMPO (60)

Chord

CHORD-TONE (C, C#, ..., B, 12), CHORD-TYPE (47), CHORD-SLASH (C, C#, ..., B, 12)

Structure

MIDLEVEL-UNIT (23), PHRASE, PART-START, PART-END (5), REP-START, REP-END (6)

To verify the efficacy of adding ***Structure***-related events, we consider 2 variants of the *Jazz Transformer*:

- **Model (A):** trained with ***Note + Metric + Chord*** events
- **Model (B):** trained with the ***complete set*** of events.

3. Getting User Feedback

In our blind listening test, users listen to 2 pieces by **Model (B)** & 2 real ones. They are asked to rate each piece in a 5-point scale on the following aspects:

- **Overall Quality (O)**
  - *Does the music sound good overall?*
- **Impression (I)**
  - *Can you recall a certain part or melody?*
- **Structureness (S)**
  - *Are there repeated motifs or phrases?*
- **Richness (R)**
  - *Do you feel the music interesting/bland?*

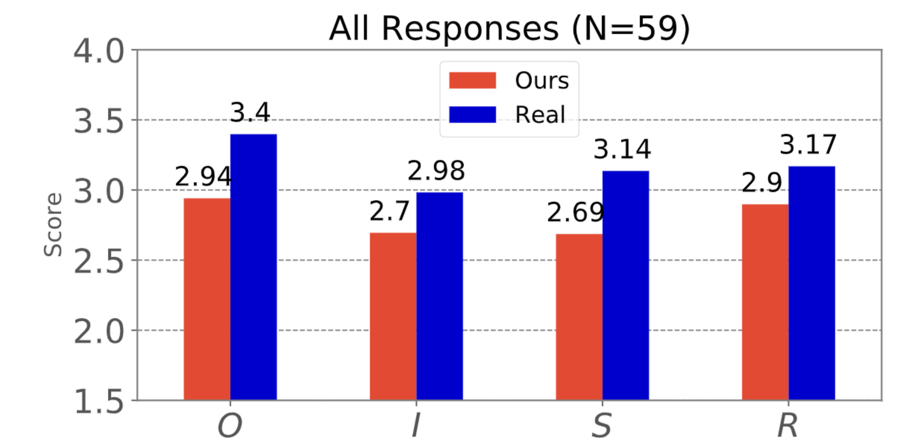

Fig. 2. *The results of user study*, showing that the ***gap*** between machine and real pieces is perceptible and statistically significant.

4. Building Objective Metrics

We develop a set of empirical measures to find out why machines still lose to humans. Furthermore, these measures can help in evaluating the models' performance even before conducting user studies.

- **Pitch Histogram Entropy:** $\mathcal{H}_1, \mathcal{H}_4$

$$\mathcal{H}(\vec{\mathbf{h}}) = -\sum_{i=0}^{11} h_i \log_2(h_i)$$

- measures instability of *pitch usage*

- **Grooving Pattern Similarity:** $\mathcal{GS}$

$$\mathcal{GS}(\vec{\mathbf{g}}^a, \vec{\mathbf{g}}^b) = 1 - \frac{1}{Q}\sum_{i=0}^{Q-1} \mathrm{XOR}(g_i^a, g_i^b)$$

- measures consistency of *rhythm*

- **Chord Progression Irregularity:** $\mathcal{CPI}$

Percentage of *unique chord trigrams*

- measures inconsistency of *harmony*

- **Structureness Indicator:** $\mathcal{SI}_3^8, \mathcal{SI}_8^{15}, \mathcal{SI}_{15}$

$$\mathcal{SI}_l^u(S) = \max_{\substack{l \le i \le u \\ 1 \le j \le N}} S$$

- examines presence of *repeated structures*

- **Continuation Prediction Challenge:** $\mathcal{C}t\mathcal{P}r$

Given an *8-bar primer*, predict the *correct 8-bar continuation* from 4 choices

- examines *overall understanding* of music.

	Model (A)		Model (B)			Real
loss	*0.80*	*0.25*	*0.80*	*0.25*	*0.10*	- -
$\mathcal{H}_1$	2.29	2.45	2.26	2.20	**2.17**	1.94
$\mathcal{H}_4$	3.12	3.05	3.04	**2.91**	2.94	2.87
$\mathcal{GS}$	**0.76**	0.69	0.75	**0.76**	**0.76**	0.86
$\mathcal{CPI}$	81.2	77.6	79.2	**72.6**	75.9	40.4
$\mathcal{SI}_3^8$	0.18	0.22	0.25	**0.27**	0.26	0.36
$\mathcal{SI}_8^{15}$	0.15	0.17	**0.18**	**0.18**	0.17	0.36
$\mathcal{SI}_{15}$	0.11	**0.14**	0.10	0.12	0.11	0.35

Table 2. *The results of objective evaluations* (numbers are the closer to **Real** the better).

Takeaways

- ***Model (B)* at 0.25 loss level is the closest competitor to humans.**
- **The models' deficiencies are manifest in:**
  - ***erraticity** of **pitch usage*** (high $\mathcal{H}_1, \mathcal{H}_4$)
  - ***lack of consistency** in **rhythm** & **harmony*** (low $\mathcal{GS}$ & high $\mathcal{CPI}$)
  - ***absence** of **longer-term structures*** (low $\mathcal{SI}_8^{15}, \mathcal{SI}_{15}$).

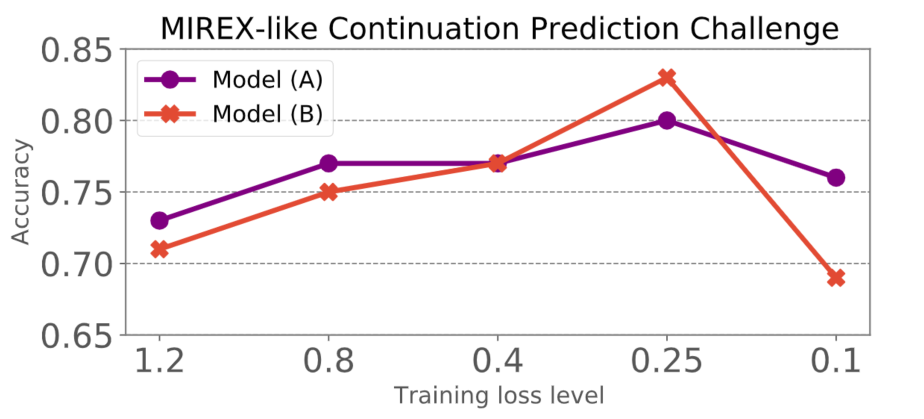

Fig. 3. *The models' performance in metric $\mathcal{C}t\mathcal{P}r$ w.r.t. loss level*, telling us that:

- **The models' knowledge of Jazz music is gained along the training process, and peaks at training loss level 0.25.**

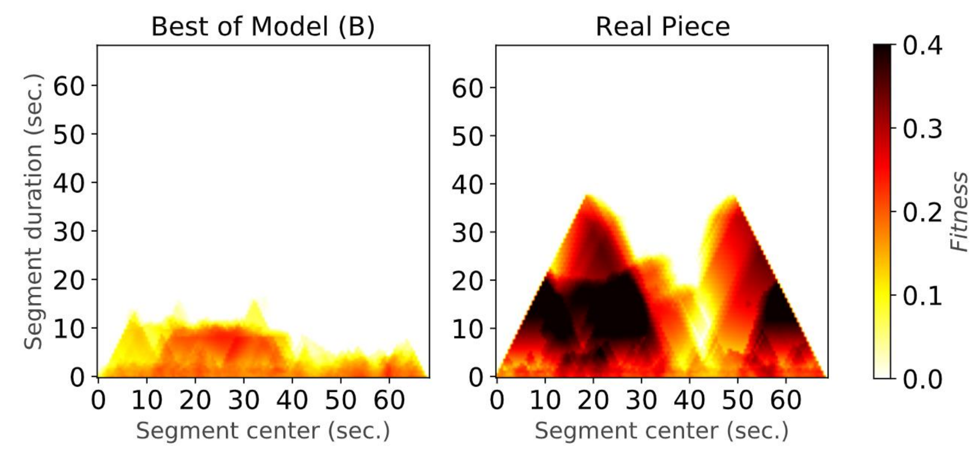

Fig. 4. *Fitness scape plots*, placing **Model (B)**'s work & a human composition in comparison. The machine composition's lack of medium- and long-term structures is clearly visualized.

5. Key Insights

- The use of ***Structure*** events **does improve** the model's compositions.
- As a composer, the *Transformer* is in fact still **far behind humans**.
- Nevertheless, its **shortcomings** are pointed out by **our objective metrics**.
- Our metrics also **shed new light on the evaluation** of machine compositions; and, **set some goals** for future work in automatic music composition **to pursue**.

bit.ly/3bDvCV5

Scan for more audio samples!

ISMIR MTL2020
National Taiwan University
AI Labs.tw 台灣人工智慧實驗室