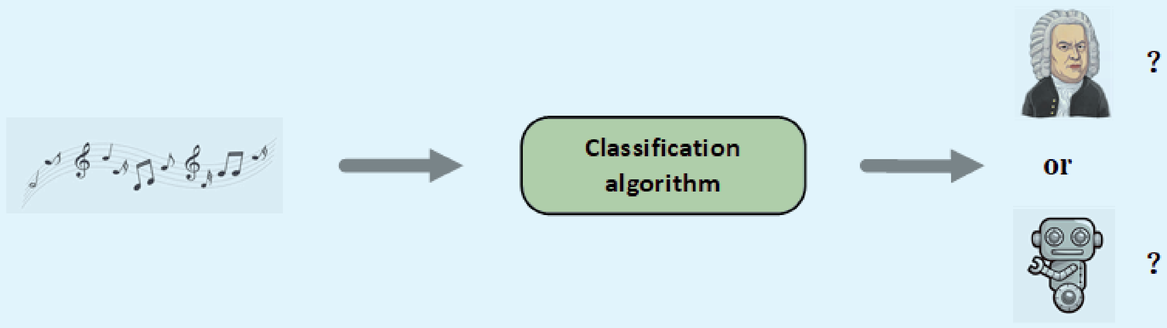
How Music Features and Musical Data Representations Affect Objective Evaluation of Music Composition: A Review of the CSMT Data Challenge 2020

Yuqiang Li[1], Shengchen Li[1], George Fazekas[2]

[1]: Xi'an Jiaotong-Liverpool University [2]: Queen Mary University of London

0. The CSMT Data Challenge 2020



The data challenge[1] required participants to develop a system that calculates the probabilities of each of the 4,000 melodies in the **evaluation** dataset[2] being composed by human composers. 6,000 fake melodies generated by a few generation models were also provided to the participants in the development dataset as auxiliary training materials.

1. Data Challenge Results

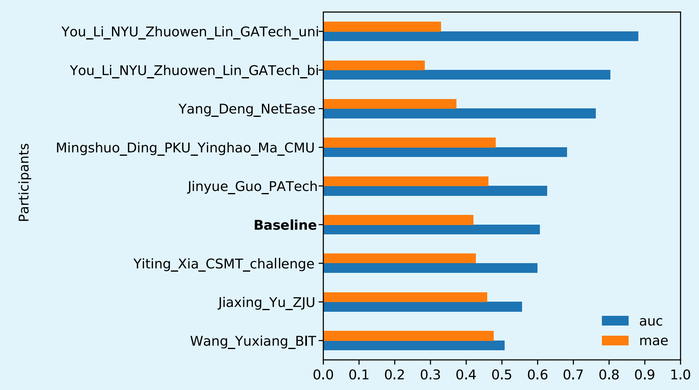


Figure 1: Model ROC AUC ranking & Model MAE

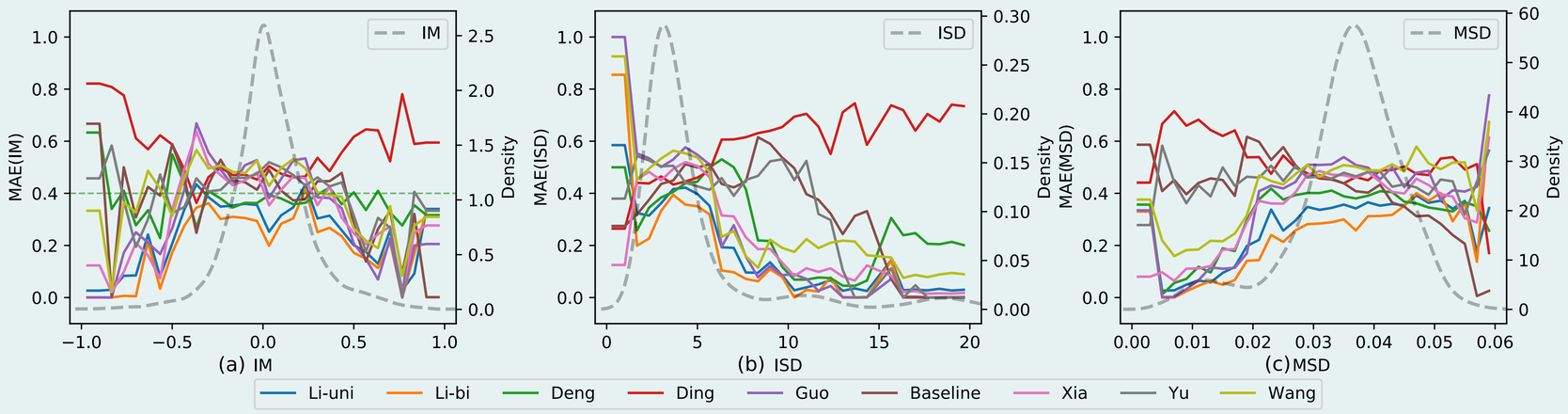
8 submissions are received according to the CSMT Data Challenge 2020 official website. System types of the submitted models:

- 6 DNN-based systems
- 1 ML system (SVM)
- 1 rule-based system

5 submissions outperformed the baseline.

The top-ranking system[3], an LSTM-based binary classifier, achieved an AUC ROC score of 0.87.

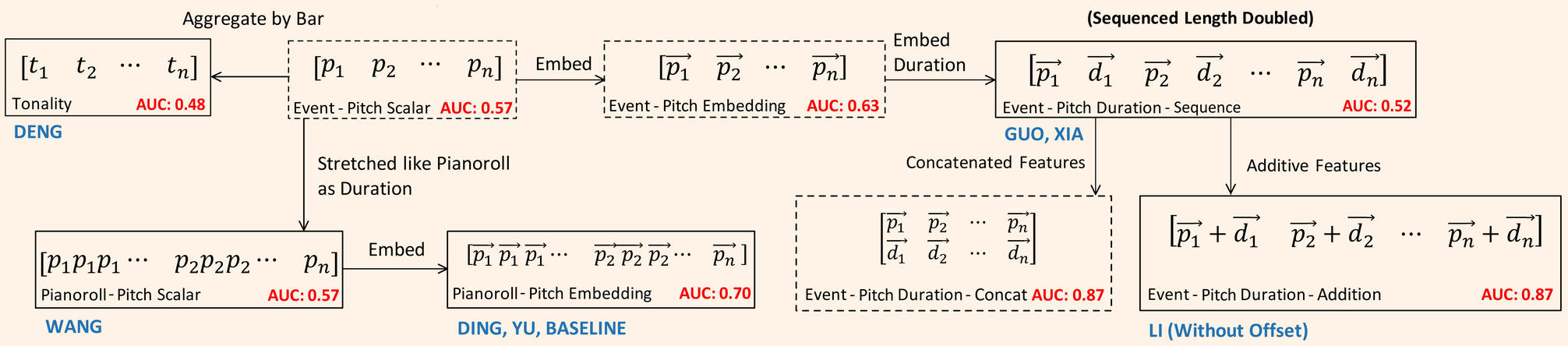
2. Model Performance & Musical Features



Interval Mean (IM), Interval Standard Deviation (ISD), and Major-scale-rate Standard Deviation (MSD) were employed to analyze the model error from different aspects.

The model MAEs vary with IM, ISD and TSD, indicating the difference in how models capture specific musical features, also the model robustness / generalization ability on extreme data.

3. Model Input Data Representation



The model input data representations used by all the submissions are plotted as boxes in this diagram, whose transformational relationships are denoted in edges.

Experiment 1: Top-ranking Model with Different Feature Selections

#	Representation	AUC
1	Event-Pitch-Scalar	0.57
2	**Tonality**	0.48
3	**Pianoroll-Pitch-Scalar**	0.57
4	Event-Pitch-Embedding	0.63
5	**Pianoroll-Pitch-Embedding**	0.70
6	**Event-PitchDuration-Sequence**	0.52
7	**Event-PitchDuration-Addition**	0.87
8	Event-PitchDuration-Concatenation	0.87

Table 1: Different representations tested on the same system. Bold ones came from the submissions.

This experiment replaced the model input data representations of the top-ranking model (LSTM binary classifiers) with all the other possible representations (as illustrated in the diagram above).

Results suggest that the **Event-PitchDuration-Concatenation** representation resulted in the best performance compared to other ones.

Experiment 2: Top-ranking Model with Different Metric Units of Time

#	Metric Unit	AUC
1	8th	0.83
2	12th	0.87
3	16th	0.88
4	24th	0.86

Table 2: Model performances using different metric units of time.

In this experiment, top-ranking system was again selected but trialed with different versions of the training dataset. The same training materials were pre-processed using different metric units of time (temporal resolutions).

As the metric unit becomes smaller, the resulted AUC ROC scores show a local maxima of 0.88, even better than the submitted version, indicating that the metric unit of time should be carefully chosen.

4. Conclusion

1. Noticeable impact of (pitch-based) music features have been observed on the model performance of objective music composition evaluation systems. These features are helpful in diagnosing the performance of music generation systems.
2. Input data representation also matters to such systems, in terms of feature selection, feature fusion, and the metric unit of time used to pre-process the musical data.

5. References

[1] "Ai composition recognition 2020," 2020. [Online]. Available: https://ai-composition-recognition2020.github.io/english.html

[2] S. Li, Y. Jing, and G. Fazekas, "A Novel Dataset for the Identification of Computer Generated Melodies in the CSMT Challenge," in *Proceedings of the 8th Conference on Sound and Music Technology.* Springer Singapore, Apr. 2021, pp. 177–186.

[3] Y. Li and Z. Lin, "Melody Classifier with Stacked-LSTM," *arXiv:2010.08123 [cs, eess]*, Nov. 2020.