Bach or Mock? A Grading Function for Chorales in the Style of J.S. Bach

Alexander Fang, Alisa Liu, Prem Seetharaman, Bryan Pardo
Northwestern University, Evanston, IL

Paper: interactiveaudiolab.github.io/assets/papers/Fang2020-MLMD.pdf Code: github.com/asdfang/constraint-transformer-bach

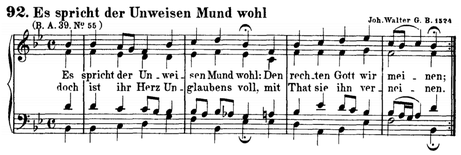
ICML International Conference On Machine Learning

MOTIVATION

Generative ML models for music creation need automatic, interpretable, and musically motivated evaluation measures of generated music.

BACH CHORALES

- Written for four parts (soprano, alto, tenor, bass) by harmonizing a Lutheran hymn
- Form a canonical dataset for music generation due to its size and stylistic consistency
- Represents a compositional challenge by balancing stylistic counterpoint rules with musical expressivity



Excerpt from BWV 308

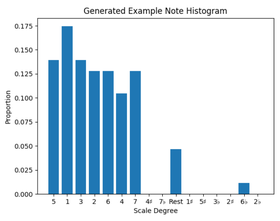
A GRADING FUNCTION FOR FOUR-PART CHORALES

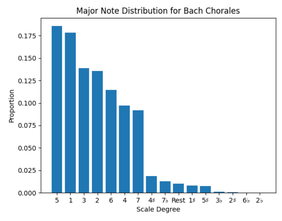
A real-valued function that evaluates quality of four-part chorales in the style of J.S. Bach along important musical features

We represent a chorale as a set of distributions, each corresponding to a musical feature.

For each feature f, use the Wasserstein metric to measure the distance between the distribution P_c^f of the given chorale c and P_{Bach}^f over the set of true Bach chorales.

$$\text{Wass}\left(P_c^f \,,\, P_{\text{Bach}}^f \right)$$





Take the weighted sum of the Wasserstein distances for the overall grade:

$$g(c) = \sum_{f \in \text{features}} w_f \cdot \text{Wass}\left(P_c^f, P_{\text{Bach}}^f\right)$$

Note that a **lower** grade represents a **better** chorale!

Table 1: Features in the grading function

Feature	Description	Musical motivation
Pitch	Distribution of pitches in scale degrees (e.g. $\hat{1}, \sharp\hat{4}, \flat\hat{5}$)	Measure overall Bach-like usage of 18th century tonality
Rhythm	Distribution of note lengths in units of quarter-notes (e.g. ♩ = 1.0, ♪ = 0.5)	Measure Bach-like usage of rhythm
{S, A, T, B} Intervals	Distribution of directed melodic interval sizes (e.g. ↑P5, ↓m2)	Evaluate musical function and contour of each voice
Harmonic qualities	Distribution of vertical harmonic qualities without root and inversion (e.g. major, minor, dominant-seventh)	Measure Bach-like usage of 18th century tonality in vertical chords
Parallel errors	Distribution of parallel fifths and octaves part-writing errors	Avoid the hallmark part-writing errors
Repeated sequence	Distribution of the length of sequences repeated in the chorale	Measure Bach-like handling of recurring motifs and intentional musical repetition

EXPERIMENTS

Show that the grading function can be used to interpret musical compositions and outperforms human experts at discrimination

We use the grading function to evaluate the output of a Transformer model with relative attention trained on Bach chorales.

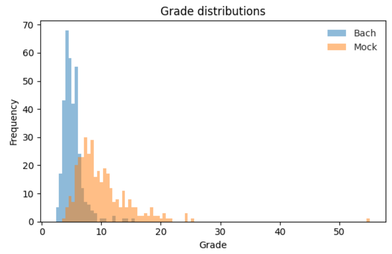


Figure 1 (left): The distribution of grades given to Bach chorales and generated chorales. The distributions are well-separated with a KS test p-value of 1e-78.

Table 2 (below): The median value for every feature in the grading function, as well as the overall grade, for Bach chorales and generated chorales. **Lower = better**.

	Note	Rhythm	Parallel Errors	Harmonic Quality	S Intervals	A Intervals	T Intervals	B Intervals	Repeated Sequence	**Overall Grade**
Bach	0.24	0.23	0.0	0.41	0.47	0.49	0.53	0.69	1.29	4.91
Mock	0.37	0.26	2.126	0.54	0.53	0.71	0.73	0.89	1.86	8.94

We can use the grading function to interpret the musical strengths and weakness of a composition!

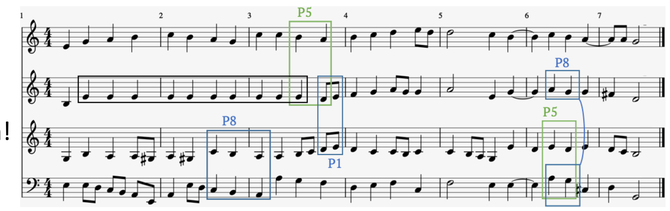


Figure 2 (above): A generated chorale receiving an overall grade of 26.0 with a parallel error distance of 5.9. P1 = parallel unison, P5 = parallel 5th, P8 = parallel 8ve.

We performed a paired discrimination test on 36 human listeners. Each pair contained one Bach and one machine-generated chorale. Selecting the chorale with the better grade results in higher accuracy than human experts at identifying Bach chorales.

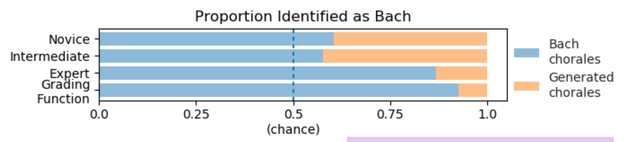


Figure 3 (above): Results of the paired discrimination experiment.

CONCLUSION

Our grading function…

- allows researchers to efficiently evaluate their models at more points during the research cycle
- sheds insight into the musical strengths and limitations of generated output
- serves as a consistent benchmark for comparing different models