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

Alexander Fang [1 2] **Alisa Liu** [1] **Prem Seetharaman** [1] **Bryan Pardo** [1]

1. Introduction

Deep generative systems that learn probabilistic models from a corpus of existing music do not explicitly encode knowledge of a musical style, compared to traditional rule-based systems (Ebciolu, 1988; Nierhaus, 2009). Thus, it can be difficult to determine whether deep models generate stylistically correct output without expert evaluation. However, human evaluation is expensive and time-consuming, limiting when it can be performed during the research cycle. Moreover, the experimental setup and execution vary greatly across human subject studies, hindering comparability of results. Therefore, there is a need for automatic, interpretable, and musically-motivated evaluation measures of generated music. Such grading functions can allow researchers to efficiently evaluate their models, shed insight into the musical strengths and limitations of generated output, and serve as a consistent benchmark for comparing different models.

In this paper, we introduce a grading function that evaluates four-part chorales in the style of J.S. Bach along important musical features. The Bach chorales represent a canonical dataset for music generation models that has been used in multiple prior works (Liang et al., 2017; Huang et al., 2017; Hadjeres et al., 2017), due to the dataset's size and stylistic consistency. We use the grading function to evaluate the output of a Transformer model, and show that the function is both interpretable and outperforms human experts at discriminating Bach chorales from model-generated ones.

2. Grading Function for Four-Part Chorales

Given a four-part chorale, our grading function[1] outputs a real-valued grade. We represent a chorale as a set of distributions, each corresponding to a musical feature important for evaluating Bach-style chorales. We implement our grading function using music21 (Cuthbert & Ariza, 2010).

For each feature f (described in Section 2.1), we use the Wasserstein metric (Rüschendorf, 1985) to measure the distance between the distribution P_c^f of the given chorale c and the ground-truth distribution P_{Bach}^f over the set of true Bach chorales. By taking a weighted sum of the Wasserstein distances over all the features (Eq. 1), we obtain the overall grade for a chorale. Note the output of the grading function is positive, and a lower grade represents a better chorale.

$$g(c) = \sum_{f \in \text{features}} w_f \cdot \text{Wass}\left(P_c^f, P_{\text{Bach}}^f\right) \tag{1}$$

2.1. Features

In this section, we describe each feature used to represent a chorale (or set of chorales). The weight $w_f = 1$ unless stated otherwise.

2.1.1. Pitch

The pitch distribution is the distribution of a chorale's pitches in scale degrees. We consider enharmonic spellings as distinct, but not octave displacements. For a concrete example, if a chorale in C Major had 60 C's, 25 F♯'s, and 15 G♭'s, the probabilities for $\hat{1}$ ("scale degree 1"), $\sharp\hat{4}$, and $\flat\hat{5}$ are .60, .25, and .15, respectively. The pitch distribution feature evaluates a Bach-like usage of tonality, distinguishing pieces that are too chromatic (e.g. twelve-tone pieces) and ones that are too stagnant (e.g. never uses *any* chromaticism).

2.1.2. Rhythm

The rhythm distribution is the distribution of note lengths in units of quarter notes, e.g. eighth notes are 0.5 units, quarter notes are 1.0. This feature serves to measure whether chorales use rhythm like Bach does: eighths and quarters as the main body, and others for decoration and variety.

2.1.3. Intervals

The interval distribution is the distribution of directed melodic interval sizes, i.e. ascending and descending intervals of the same distance are different. Each voice (soprano, alto, tenor, bass) serves a different musical function; specifically, melodies in soprano parts have the most intervallic

[1]Department of Computer Science [2]Bienen School of Music, Northwestern University, Evanston, IL, USA. Correspondence to: Alexander Fang <alexanderfang2019@u.northwestern.edu>, Alisa Liu <alisa@u.northwestern.edu>, Prem Seetharaman <prem@u.northwestern.edu>, Bryan Pardo <pardo@northwestern.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

[1]https://github.com/asdfang/constraint-transformer-bach/tree/master/Grader

Table 1. The median value (standard deviation) for every feature in the grading function, as well as the overall grade, for Bach chorales and generated chorales. Lower values represent better chorales. We can see that the model struggles with avoiding parallelisms.

	Note	Rhythm	Parallel Errors	Harmonic Quality	S Intervals	A Intervals	T Intervals	B Intervals	Repeated Sequence	**Overall Grade**
Bach	0.24 (0.15)	0.23 (0.14)	0.0 (0.69)	0.41 (0.2)	0.47 (0.28)	0.49 (0.23)	0.53 (0.24)	0.69 (0.4)	1.29 (0.88)	4.91 (1.63)
Generated	0.37 (0.22)	0.26 (0.14)	2.16 (3.22)	0.54 (0.31)	0.53 (0.35)	0.71 (0.34)	0.73 (0.38)	0.89 (0.68)	1.86 (2.81)	8.94 (4.64)

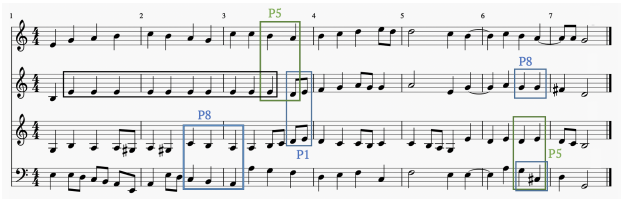

Figure 1. A generated chorale receiving an overall grade of 26.0 with a parallel error distance of 13.9 and repeated sequence distance of 5.9. The features with the largest values represent weaknesses of the composition.

Figure 2. Results of the paired discrimination experiment carried out on human listeners. The grading function "picks" the chorale that receives the better grade and achieves 92% accuracy, outperforming human experts at 86.7%.

variety, bass parts leap more frequently for harmony, and tenor and alto parts tend to employ mostly small intervals. Therefore, we measure the interval distribution separately for each voice, for a total of four interval distributions.

2.1.4. Harmonic Qualities

The harmonic qualities distribution describes the usage of vertical harmony by keeping only the quality, e.g. "D Major" would be reduced to "major." This feature also helps encourage a Bach-like usage of 18th century tonality by majority of major, minor, and dominant-seventh chords.

2.1.5. Parallel Errors

The parallel errors distribution is the distribution of occurrences of the hallmark part-writing errors: parallel fifths and octaves (including unisons) in similar and contrary motion. Observe that what matters is not only the distribution between parallel fifths and octaves, but also the count of these errors relative to the total number of notes. Therefore, the Wasserstein distance for this feature is multiplied by $w_{\text{parallel errors}} = \frac{\text{error to note ratio of chorale } c}{\text{error to note ratio of Bach}}$. This weight is large if the given chorale has a large error to note ratio compared to real Bach chorales, thereby penalizing the chorale.

2.1.6. Repeated Sequences

The repeated sequence distribution is the distribution of the length (in units of quarter notes) of sequences containing at least two notes and appearing at least twice in the chorale, in order to promote a Bach-like handling of recurring motifs and intentional musical repetition. To identify repeated sequences, we use the dynamic programming algorithm in (Hsu et al., 1998).

3. Experiments

We now show that the grading function provides interpretable output and is a promising substitute for human evaluation. We used the grading function to evaluate the output of a Transformer model (Vaswani et al., 2017) with relative attention (Huang et al., 2018) trained on a corpus of 351 Bach chorales, using the same data representation as in (Hadjeres et al., 2017).

The grade distribution for Bach chorales and generated chorales is very well-separated with a KolmogorovSmirnov test p-value of $1e{-}78$. In Table 1, we compare the median value of every feature in the grading function. Generated chorales do worse than Bach chorales in every feature.

To further show the grading function's interpretability, we display a badly graded generated chorale in Figure 1. We see especially large distances for its parallel error and repeated sequence features. Indeed, the grading function automatically found six total parallel errors and identified an abnormally long sequence of repeated quarter notes (the repeated E's in measures 1–3 of the alto voice).

To compare our grading function to human performance, we performed a paired discrimination test with $n = 36$ responses. We assessed the musical expertise of our participants through a series of pre-test questions, and assigned them to one of three groups: novice ($n_0 = 16$), intermediate ($n_1 = 15$), and expert ($n_2 = 5$). In the paired discrimination test, we presented three *pairs* of audio examples representing complete chorales, one Bach and one generated, and asked participants to select the one composed by Bach. In Figure 2, we compare the human pick to selecting the chorale that receives a better grade. We find that the grading function achieves 92.6% accuracy, outperforming human experts at 86.7%.

References

Allan, M. and Williams, C. K. I. Harmonising chorales by probabilistic inference. In *International Conference on Neural Information Processing Systems (NIPS)*, 2004.

Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. 2012.

Cuthbert, M. S. and Ariza, C. music21: A toolkit for computer-aided musicology and symbolic music data. In *Conference of the International Society of Music Information Retrieval (ISMIR)*, pp. 637–642, 2010.

Ebciolu, K. An expert system for harmonizing four-part chorales. *Computer Music Journal*, 12, 1988.

Hadjeres, G., Pachet, F., and Nielsen, F. DeepBach: a steerable model for Bach chorales generation. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1362–1371, 2017.

Hsu, J.-L., Chen, A. L. P., and Liu, C.-C. Efficient repeating pattern finding in music databases. In *International Conference on Information and Knowledge Management*, pp. 281288, New York, NY, USA, 1998. Association for Computing Machinery.

Huang, C. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Hawthorne, C., Dai, A. M., Hoffman, M. D., and Eck, D. An improved relative self-attention mechanism for transformer with application to music generation. *CoRR*, abs/1809.04281, 2018.

Huang, C.-Z. A., Cooijmans, T., Roberts, A., Courville, A., and Eck, D. Counterpoint by convolution. In *Conference of the International Society of Music Information Retrieval (ISMIR)*, 2017.

Liang, F. T., Gotham, M., Johnson, M., and Shotton, J. Automatic stylistic composition of bach chorales with deep lstm. In *Conference of the International Society of Music Information Retrieval (ISMIR)*, 2017.

Nierhaus, G. *Algorithmic Composition: Paradigms of Automated Music Generation*. Mathematics and Statistics. Springer Vienna, 2009. ISBN 9783211755402.

Rüschendorf, L. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70 (1):117–129, 1985.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, 2017.