Computational Linguistics Metrics for the Evaluation of Two-Part Counterpoint Generated with Neural Machine Translation

Stefano Kalonaris
RIKEN AIP
Japan

Thomas McLachlan
RIKEN AIP
Japan

Anna Aljanaki
University of Tartu
Estonia

Abstract

In this paper, two-part music counterpoint is modelled as a neural machine translation (NMT) task, and the relevance of automatic metrics to human-targeted evaluation is investigated. To this end, we propose a novel metric and conduct a user study comparing it to the automatic scores of a base model known to perform well on language tasks along with different models obtained with hyper-parameter tuning. Insights of this investigation are then speculatively extended to the evaluation of generative music systems in general, which still lacks a standardised procedure and consensus.

1 Introduction

The modelling and generation of contrapuntal music has been tackled using a plethora of approaches, ranging from rule and constraint-based (Ebcioğlu, 1988; Tsang and Aitken, 1991) to grammars (Gilbert and Conklin, 2007; Quick and Hudak, 2013), statistical methods such as Hidden Markov Models (Farbood and Schöner, 2001; Allan and Williams, 2004), combinations of the latter with pattern-matching models (Cope, 1992) or templates (Padilla and Conklin, 2018), and neural networks. Among the latter, generative adversarial networks (GAN) (Dong et al., 2018), variational autoencoders (VAE) (Roberts et al., 2017), and convolutional neural networks (CNN) (Huang et al., 2017) have proven successful. Recurrent neural networks (RNN), particularly long short-term memory (LSTM) architectures (Sturm, 2018; Simon and Oore, 2017), and more recent attention-based models are also increasingly applied for the generation of music (Payne, 2019; Huang et al., 2018; Hawthorne et al., 2018); however, language-based models have not been employed as much for modelling contrapuntal music.

In a recent study (Nichols et al., 2021), two-part counterpoint generation was treated as a NMT task, by considering one part as the source language, and the other as the target language (see Figure 1). We extend the NMT analogy from the formulation of the task to the evaluation of the system's musical output, and consider standard metrics used in translation. A novel variation of a human-based metric is proposed and compared to automatic metrics via a user study, and inter-annotator agreement is also assessed. This paper's contribution can be summarised as 1) a novel application of computational linguistics methods for the evaluation of counterpoint generated using NMT and 2) reusable insights in the broader domain of generative music systems.

2 Data

We used the Multitrack Contrapuntal Music Archive[1] (MCMA) as the training corpus, comprising only track-separated contrapuntal pieces, each ranging from two to six tracks. The dataset of source-target musical sentences for training our model(s) was obtained by making all $\binom{k_i}{2}$ combinations of pairs of tracks, where i indexes the works. This yielded $1,418$ track pairs, which were then segmented into $17,734$ non-overlapping four-bar chunks. No data augmentation was performed. Instead, all pieces were normalised to a key with zero flats/sharps (notably, $C/Amin$). Events in each score segment were encoded similarly to (Nichols et al., 2021) although we did not require strictly monophonic voices[2], and we relied on the model's inbuilt positional encoding (see Section 3), thus omitting a *beat position* token.

3 Model & Tuning

In this work we have used the *Transformer* model (Vaswani et al., 2017) which allows each

[1] https://mcma.readthedocs.io/en/latest/contents.html

[2] splits were encoded as *Chord* objects.

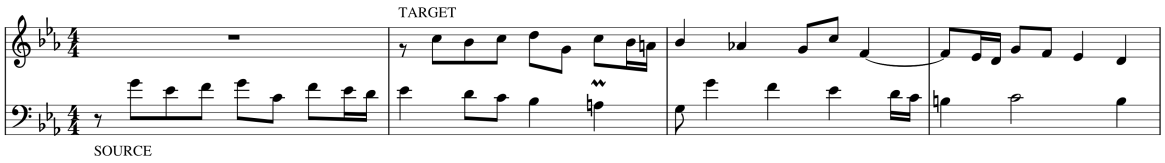

Figure 1: First four bars of the top two voices in J.S. Bach's Fugue in C minor, BWV 871, from *The Well-Tempered Clavier*, viewed as a NMT compositional task.

position in a sequence to attend (Bahdanau et al., 2015) to any other position. An Encoder or Decoder layer has *Self-Attention*, to attend to any position on the layer input, and the Decoder layer has an additional *Encoder-Decoder Attention* in which each decoding step can attend to all the Encoder final layer outputs. For the finer details we refer the reader to the original paper, aforementioned. There, the authors had success in translation with $d = 512$, $l = 6$, $n_h = 8$, which we refer to as the *Base* model (where d is the dimension of each layer and embedding, l the number of layers, and n_h is the number of attention heads).

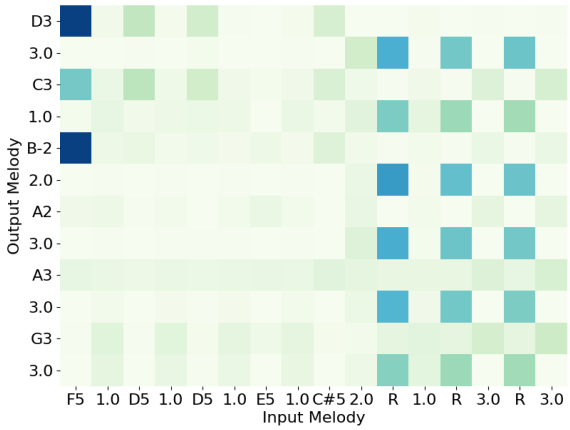

Figure 2: Example of an attention map to the encoder outputs while decoding a short validation phrase. The output is read from top to bottom. (Layer 3, $head_1$)

However, in music generation there are several differences from the language setting, such as having a shared vocabulary between input and output; a reduced vocabulary of 190 tokens compared to the ten thousands to millions of words in language; and considerable differences in the expectation of grammar. In language translation, *attention scores* to the input positions during decoding normally reveal a strong attention along the diagonal. In our music implementation, this is less apparent; however in Figure 2 we see the model attend to positions of rest in the source melody when considering note durations, and to the first few source pitches when beginning generation of a melody.

These differences motivated the re-tuning of some of the main parameters of the Transformer, namely $l = [2, 4, 6, 8]$, $d = [256, 512, 1024]$), and $n_h = [4, 8, 16]$. In addition, since it was not clear what criteria should be used to halt the training, each model was trained for a set number of epochs ($n_e = [6, 8, 10, 12, 20]$). These epoch ranges were determined by observing the over-fitting behaviour of the base model (up to 60 epochs). This formed 180 parameter sets, which were trained simultaneously on 180 NVIDIA V100 GPUs on *RAIDEN* at *RIKEN* within 24 hours. The question still remained on how these trained models should be evaluated, and how a best model could be selected.

4 Evaluation

In an effort to balance the need for formative assessments aimed at establishing reliable objective measures (Yang and Lerch, 2018) with the necessity to put generative music output back to the domain experts realm (Sturm and Ben-Tal, 2017), we investigated both automatic and human-targeted metrics.

4.1 Automatic metrics

Of the common automatic metrics for translation tasks, we used Loss, Token Accuracy, Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), Perplexity (Brown et al., 1992), and Word Error Rate (WER) (Klakow and Peters, 2002).

Table 1 shows a few selected candidates favoured by 1 or 2 automatic metrics during tuning. Interestingly, with the exception of *BestPPL*, smaller models were favoured ($d = 256$). The generation quality seemed to be comparable across the selected models, given the amount of data we trained on. So these metrics perform at a similar level when applied to music.

Model	d	l	n_h	n_e	Loss	Acc	WER	BLEU	ROUGE	PPL
Base	512	6	8	10	1.057	0.681	48.16	51.31	74.77	2.904
AccBLEU	256	4	4	20	1.040	**0.689**	43.53	**53.95**	76.06	2.862
LossROUGE	256	6	16	8(10)	1.028 (**1.022**)	0.681	63.54	47.43	**76.68**	2.812
BestWER	256	2	8	12	1.033	0.682	**38.27**	53.74	74.53	2.833
BestPPL	512	8	8	8	1.022	0.685	57.03	50.99	75.45	**2.798**

Table 1: Candidate models selected by 1 or 2 metrics on the validation set. *Acc* refers to Token Accuracy, and *ROUGE* refers to the ROUGE-1 F1 score. The numbers in brackets come from the Loss variant.

Model	NC		NLTM	
	KLD	OA	KLD	OA
Base	0.0026	0.9380	0.0266	**0.9696**
AccBLEU	**0.0008**	**0.9549**	0.0107	0.9540
LossROUGE	0.0011	0.9458	0.0166	0.9498
BestWER	0.0011	0.9499	0.0162	0.9517
BestPPL	0.0017	0.9430	**0.0104**	0.9576

Table 2: Kullback–Leibler Divergence (KLD) and Overlapping Area (OA) between the models' dataset intra-set PDF and the inter-set PDF. Shown for total notes used (NC) and note length transition matrix (NLTM).

As a baseline, we used Yang and Lerch's (2018) evaluation method. Their exhaustive cross-validation based on intra and inter-test measurements, and on Kullback–Leibler Divergence (KLD) and Overlap Area (OA) (see Table 2) also failed to single out a best model.

To get a better understanding of how automatic metrics correlate to music generation quality, we considered human evaluation.

4.2 Human-targeted metrics

Rather than relying on Turing-type tests, which have been sufficiently criticised in (Ariza, 2009), we consider instead the *human-targeted translation edit rate* (HTER) (Snover et al., 2006) and propose a variant that can be used in the music domain. In HTER, typically, human annotators generate a new targeted reference by editing the machine generated target (hypothesis) until it has the same meaning as an original reference translation. Subsequently the *translation edit rate* (TER) is calculated between the new targeted reference and the machine hypothesis.

4.2.1 Music and Semantics

A problem with using HTER 'as is', resides in the contentious issue of whether music, as opposed to language, is semantic or not. There seems to be a general consensus toward the latter, despite studies (Koelsch et al., 2004) showing the ability of music excerpts to prime words. Psychoacoustic and socio-cultural specific properties of music might be able to induce emotion or infer meaning (Meyer, 1956), and there is growing interest in musical semantics (Schlenker, 2017) which, in turn, draw from *Gestalt* theory-based approaches to music (Lerdahl and Jackendoff, 1982). However, it remains unclear how would annotators edit the hypothesis melody so as to have the same "meaning" of the reference target melody. Because of this issue on semantics we consider a simple variation on HTER.

4.2.2 HER

We propose a new metric inspired by the HTER, whereby annotators, all domain experts, are not provided with the (original) reference. Instead, they are asked to edit the generated hypothesis directly until it is, in their domain expertise, sufficiently acceptable as a musical complement/response to the source melody (see Figure 3 for an example). Then, a suitable distance metric (we used the WER) between the obtained targeted reference and the generated hypothesis is calculated. We call this metric, simply, human-targeted edit rate (HER).

5 Experiment

We generated musical mini-scores from the test set for the base model and for the models with the highest validation score on the automatic metrics described in Section 4.1 (AccBLEU, LossROUGE, BestPPL, and BestWER models). The test set models' targets produced $3,289$ mini-scores (each between 2 and 6 bars in length, approximately) for each model. Of these, approximately half (the percentage varied depending on the specific model) were filtered out for

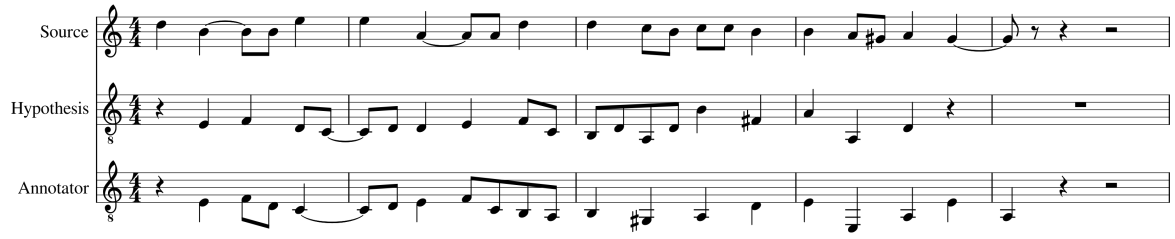

Figure 3: Example of a targeted reference obtained from editing the hypothesis.

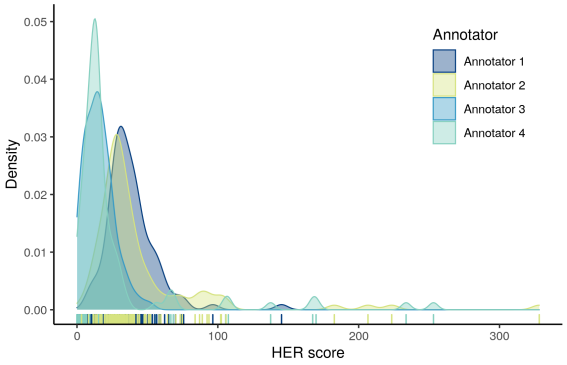

Figure 4: Distributions of the HER scores by annotator displayed as KDE for clarity purposes, as these are easier to visually process than overlapping histograms.

not having an end-of-sequence token, for being badly formatted (not alternating correctly between chord/note/rest and duration tokens) or for having less than three notes in any given part. Subsequently, $1,067$ matching mini-scores (across all models) were identified, and 100 of these (20 per model) were randomly selected to be given to 4 annotators. The LossROUGE model scored the lowest mean HER (29.69 ± 21.85). We calculated inter-annotator agreement using the Krippendorff's alpha coefficient (Krippendorff, 2004) and the intraclass correlation coefficient (ICC) for a fixed set of annotators rating each target (Bartko, 1966). We note that LossROUGE is the least reliable in terms of agreement. The rest of the models range between poor to moderate agreement. The overall inter-model agreement stood at Krippendorff's alpha of 0.388 and ICC of 0.411. The average amount of edits varied by annotator according to their personal error tolerance, creating a variability in HER. After normalisation, we obtain Kripperndorff's alpha of 0.483 and ICC of 0.61. These results are summarised in Table 3 and in Figure 4.

Model	**Mean±Std**	**Kr.** α	**ICC**
Base	31.45 ± 35.53	**0.493**	0.431
AccBLEU	35.54 ± 46.2	0.410	**0.452**
LossROUGE	**29.69±21.85**	0.164	0.260
BestWER	32.75 ± 37.20	0.306	0.374
BestPPL	30.54 ± 25.38	**0.493**	0.322

Table 3: Intra-model HER scores and agreement.

6 Conclusion & Reusable Insights

We presented a study on computational linguistic metrics applied to the evaluation of two-part music counterpoint generated with a language-based model. A novel human-targeted metric (HER) was proposed, to correlate automatic translation metrics to human judgement, in the music domain. The HER metric bypasses the contentious notion of human/machine discrimination and, while subjectivity is still part of the process (no two annotators would edit the generated hypothesis in an identical way), it does not require defining musical features of interest in advance. It is, instead, assumed that domain practitioners have their own definitions of musical fitness and edit the model's output accordingly, and that individual biases can be measured via inter-annotator reliability.

In our study, we hoped that the HER score would help elucidate the strength of NLP automatic translation metrics for music generation. While this study proved inconclusive given the low inter-annotator agreement, also reported in other music annotation tasks (Gjerdingen and Perrott, 2008; Flexer and Grill, 2016; Koops et al., 2019), it nevertheless provides an original approach which can be employed to evaluate other generative music systems.

References

Moray Allan and Christopher K. I. Williams. 2004. Harmonising Chorales by Probabilistic Inference. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04.

Christopher Ariza. 2009. The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems. *Computer Music Journal*, 33(2):48–70.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

John J. Bartko. 1966. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports*, 19(1):3–11.

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics*, 18(1):31–40.

David Cope. 1992. Computer Modeling of Musical Intelligence in EMI. *Computer Music Journal*, 16(2):69–83.

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*.

Kemal Ebcioğlu. 1988. An Expert System for Harmonizing Four-Part Chorales. *Computer Music Journal*, 12(3):43–51.

Mary Farbood and Bernd Schöner. 2001. Analysis and Synthesis of Palestrina-Style Counterpoint Using Markov Chains. In *Proceedings of International Computer Music Conference*, ICMC'01.

Arthur Flexer and Thomas Grill. 2016. The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. *Journal of New Music Research*, 45(3):239–251. PMID: 28190932.

Édouard Gilbert and Darrell Conklin. 2007. A Probabilistic Context-Free Grammar for Melodic Reduction. In *International Workshop on Artificial Intelligence and Music, The Twentieth International Joint Conference on Artificial Intelligence*, IJCAI'07.

Robert O. Gjerdingen and David Perrott. 2008. Scanning the Dial: The Rapid Recognition of Music Genres. *Journal of New Music Research*, 37(2):93–100.

Curtis Hawthorne, Anna Huang, Daphne Ippolito, and Douglas Eck. 2018. Transformer-NADE for Piano Performances. In *NIPS 2018 Workshop on Machine Learning for Creativity and Design*.

Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck. 2017. Counterpoint by Convolution. In *International Society for Music Information Retrieval (ISMIR)*.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, and Douglas Eck. 2018. Music Transformer: Generating Music with Long-Term Structure. *arXiv preprint arXiv:1809.04281*.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Commun.*, 38(1):19–28.

Stefan Koelsch, Elisabeth Kasper, Daniela Sammler, Katrin Schulze, Thomas Gunter, and Angela D. Friederici. 2004. Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience*, 7(3):302–307.

Hendrik Vincent Koops, W. Bas de Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller, and Anja Volk. 2019. Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, 48(3):232–252.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage.

Fred Lerdahl and Ray Jackendoff. 1982. *A generative theory of tonal music*. MIT Press, Cambridge, MA.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Leonard B. Meyer. 1956. *Emotion and meaning in music*. University of Chicago Press, Chicago.

Eric P. Nichols, Stefano Kalonaris, Gianluca Micchi, and Anna Aljanaki. 2021. Modeling Baroque Two-Part Counterpoint with Neural Machine Translation. In *Proceedings of the International Computer Music Conference*, Santiago, Chile. International Computer Music Association. Preprint available: https://arxiv.org/abs/2006.14221.

Victor Padilla and Darrell Conklin. 2018. Generation of Two-Voice Imitative Counterpoint from Statistical Models. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(3):22–32.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Christine Payne. 2019. MuseNet. https://openai.com/blog/musenet.

Donya Quick and Paul Hudak. 2013. Grammar-based Automated Music Composition in Haskell. In *Proceedings of the First ACM SIGPLAN Workshop on Functional Art, Music, Modeling & Design*, FARM'13, pages 59–70, New York, NY, USA. ACM.

Adam Roberts, Jesse Engel, and Douglas Eck. 2017. Hierarchical Variational Autoencoders for Music. In *NIPS 2017 Workshop on Machine Learning for Creativity and Design*.

Philippe Schlenker. 2017. Outline of Music Semantics. *Music Perception*, 35(1):3–37.

Ian Simon and Sageev Oore. 2017. Performance RNN: Generating Music with Expressive Timing and Dynamics. https://magenta.tensorflow.org/performance-rnn.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Translation in the Americas (AMTA)*.

Bob Sturm. 2018. What do these 5,599,881 parameters mean? : An analysis of a specific LSTM music transcription model, starting with the 70,281 parameters of its softmax layer. In *Proceedings of the 6th International Workshop on Musical Metacreation (MUME 2018) :*. QC 20181106.

Bob Sturm and Oded Ben-Tal. 2017. Taking the Models back to Music Practice: Evaluating Generative Transcription Models built using Deep Learning. *Journal of Creative Music Systems*, 2(1).

Chi Ping Tsang and M. Aitken. 1991. Harmonizing Music as a Discipline in Constraint Logic Programming. In *Proceedings of the International Computer Music Conference*, ICMC'91, pages 61–64.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Li-Chia Yang and Alexander Lerch. 2018. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784.