Meet HER: A Language-based Approach to Generative Music Systems Evaluation

Stefano Kalonaris[1] and Anna Aljanaki[2]

[1] RIKEN AIP, Japan
[2] University of Tartu, Estonia

Abstract. Consensus and standardised procedures in the evaluation of generative music systems are hard to come by. In this paper, a novel human-based method adapted from a machine translation metric is proposed and argued to be useful for direct comparison between different systems. To this end, detailed results from a study using this metric for the evaluation of a language-based model for generative counterpoint is presented.

Keywords: evaluation metrics, computational linguistics, language models for music generation

1 Introduction

The evaluation of generative music systems' output is a complex task, having to account for disparate concerns and needs, and involving considerations of musicological, mathematical, psychological, and aesthetic nature. Thus, it is not surprising that different methods are advocated, and that, to date, little consensus has been reached regarding standardised procedures.

In Yang & Lerch (2018), a comprehensive evaluation procedure (hereinafter, *MGEVAL*) is proposed, whereby both absolute and relative metrics are obtained from the analysis of 5 pitch-based features and 4 rhythm-based features. The method is then tested for dataset evaluation, system comparison, and performance evaluation. The features used are mostly frequency distributions and averages, which can be problematic when time is a paramount factor of the phenomenon under consideration. In the same paper is, however, conceded that human evaluation remains preferable when assessing music generation, despite the many problems that it involves.

A common strategy in this context is using the so-called musical Turing tests. These tests normally involve determining, out of a choice of two outputs, which is computer generated and which is human composed. They have been criticised in that they are more akin to "discrimination tests" and/or Musical Output Toy Tests (MOtT), rather than Turing tests (Ariza, 2009), and for their inability to shed light on the intelligence of the system or the music quality of its output.

Aware of the dangers of musical Turing tests, but asserting the importance of re-contextualising the generative music systems' output with a firm grounding

in domain practice, Sturm & Ben-Tal (2017) combine statistical analysis with several other human-based evaluations, to include musicological analysis, testing the music knowledge limits of the model, assessing the usefulness of the model in an aided-composition scenario, and finally asking domain practitioners how well the generated output integrates with their personal practice and experience.

This paper presents our contribution to this discourse, which also foregrounds the paramount role of human-based evaluation, and strives to integrate it with objective and automatic metrics, in one generalisable procedure. The latter was originally proposed in (Kalonaris, McLachlan, & Aljanaki, 2020), in the context of modeling the composition of two-part polyphony as a machine translation problem of generating a musical part (target) given another (source). However, in this paper, we 1) expand on the formal definition of the method, 2) provide detailed insight and examples of the results in the aforementioned study, and 3) present a more extensive discussion regarding existing feature-based approaches, by using MGEVAL as a baseline.

2 HER

The *human-targeted edit rate* (HER) is a variation on the human-targeted translation edit rate (HTER) (Snover, Dorr, Schwartz, Micciulla, & Weischedel, 2006). In HTER, human annotators are given a source sentence, the machine generated translation (hypothesis), and one or more reference translations. Subsequently, they edit the hypothesis until it has the same meaning of one of the references. Then, the translation error rate (TER) is calculated by normalising the number of edits that were applied by the average word length of the references (see Equation 1).

$$TER = \frac{\#\text{ edits}}{\text{average }\#\text{ of reference words}} \tag{1}$$

There are good reasons why a direct porting of HTER to the music domain would not be feasible. For example, linguistic and musical meanings are, fundamentally, unlike each other, and it is improbable that different music can convey a common set of meanings (Becker, 1986). It would be challenging to interpret what the meaning of a musical sentence is, and even harder to judge/establish an equivalence with another musical sentence, at this level. Arguably, "meaning" in this context would be a combination of musical syntax (*e.g.*, metrical structure, form, etc.), semantics (*e.g.*, psychological and/or affective properties of the excerpt) and pragmatics (*e.g.*, inferring the original genre/style from the excerpt, thus deducing appropriate musical idiom for the task), but its univocal definition beyond subjective interpretations would be impractical, if attainable.

As for the reference sentence, this could be simply the original counterpart to the source from a given piece excerpt. But it would not be feasible, or at least expensive and time consuming, to produce more references for the annotators. That is, these would have to be composed *ad hoc*, unless the excerpt referred to a piece with several voices.

To address these problems, we propose the following procedure, instead.

1. **Referenced-target**: annotators are given the source sentence and the hypothesis. Based on their domain expertise, they edit directly the hypothesis until it is sufficiently appropriate as a musical complement to the source. This is, of course, a subjective judgement, however, annotators are specifically instructed to apply the minimum number of necessary edits that guarantee them a satisfactory musical result. An illustration of this procedure is shown in Figure 1.

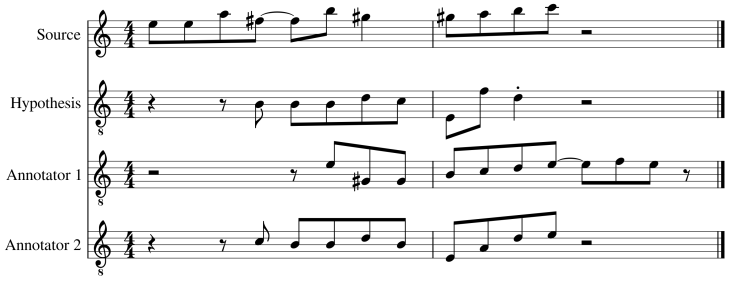

Fig. 1: From the hypothesis to a referenced-target, shown for two annotators.

2. **HER score**: the so obtained referenced-target is then compared to the hypothesis, by means of some edit distance measure. Options for such a metric are: the Word Error Rate (WER) (Klakow & Peters, 2002), the Levenshtein distance (Navarro, 2001), or the Mongeau-Sankoff algorithm (Mongeau & Sankoff, 1990), among others. Optionally, these scores could be normalised over all annotators, as this would allow direct comparison between different systems or tasks. Aggregate functions can then be used to report scores for the system being evaluated.
3. **Inter-annotator agreement**: to complete the procedure, inter-annotator agreement (IAA) tests are performed. Options include tests from the *Kappa* and *Alpha* family (Artstein, 2017) and correlation coefficients. It is worth noting that inter-annotator disagreement in the music domain when compared to linguistics might be more pronounced. The subjectivity of music perception has been noted in classification (Gjerdingen & Perrott, 2008), melodic similarity (Flexer & Grill, 2016), and harmonic annotation tasks (Koops et al., 2019). Further disparity can be introduced by the dependence on specific music-theoretical axioms (*e.g.*, Riemannian vs. Schenkerian analytical frameworks) (Selway et al., 2020), assumptions on style and/or genre, and so forth.

2.1 General Case

Besides the translation paradigm, it is possible to generalise the HER procedure by simply removing the source element altogether. In this broader scenario, annotators are given the generative music output from the system, and they proceed similarly in editing it until it is sufficient and satisfying, according to their domain expertise. As a general objective of this evaluation method, we concede that HER is more concerned with *acceptability* and *quality*, rather than *creativity*, thus falling in the "weak" objectives category, according to Ritchie (2019).

2.2 Advantages and Drawbacks

The HER procedure, unlike MGEVAL, is not limited to corpus-based systems, nor it is bound to a fixed bar length input and does not impose monophonic constraints on the source. HER does not engage in the comparison between generated and real musical reference, thus avoiding human/machine discrimination altogether. It does, however, integrate automatic evaluation and domain expertise. HER explicitly requires the latter on behalf of the annotators. Because of this, it is not necessary to define musical features of interest in advance. Instead, HER relies on the annotators' musical competence to inform their edits. That doesn't mean that HER is devoid of predefined notions (*e.g.*, choices regarding the distance metric, music representation, etc.) or of individual biases. The latter, however, can be smoothed thanks to averaging, and controlled by inter-annotator reliability tests. HER was conceived in the context of symbolic music, and it would be arduous to apply it to generative music systems with raw audio output, unless intermediate processing (such as automatic music transcription) was applied. Furthermore, this method also implicitly assumes that the annotators are digitally apt to use a score editor, to apply the desired edits to the musical scores and save these back in the required format. In this sense, it might make the sourcing and training of the annotators more labour-intensive. This, combined with the domain expertise constraint, can limit HER's applicability to large-scale evaluation exercises.

Currently, HER requires further experimentation. For example, producing well-designed instructions for the annotators to ensure comparable strategies and/or aims can be challenging and this should be considered more formally in the future. Notwithstanding the current limitations, as a proof of concept, we now present in detail the results of the first study in which it was used.

3 Experimental Results

In (Kalonaris et al., 2020), similarly to (Nichols, Kalonaris, Micchi, & Aljanaki, 2021), an attention-based model (Vaswani et al., 2017), hereinafter referred to as *base* model, was employed. There are, however, several differences between the two experiments, particularly regarding the corpus choice and the data encoding

(please refer to the aforementioned papers, for more details). Furthermore, in Kalonaris et al. (2020), hyper-parameter optimisation was also performed, and models which favoured 1 or 2 of automatic metrics[3] common in natural language translation were then selected, along with the base model. Hereinafter, we refer to these models as *AccBLEU*, *LossROUGE*, *BestWER*, and *BestPPL*. All these seemed to produce comparable results, and it was not possible to select a best performing model on the basis of the automatic metrics, alone. This motivated the ideation and application of HER, to correlate automatic metrics to human judgement, and 20 (matching) mini-scores (between 2 and 6 bars long) for each model were randomly selected to be given to annotators.

3.1 Crowdsourcing the annotations

We used the Yandex.Toloka crowdsourcing platform[4] to gather annotations. Ensuring high-quality standards is the biggest concern when working with crowdsourced data. In this study's call for participation, people with high education in music were invited to participate in a test with questions in harmony and music theory. The successful participants were invited for a second test, where they had to correct a sample score according to HER guidelines. Two online and two offline participants (mean age=41.75, all four male) were recruited. Participants were paid a fair fee for each task.

3.2 HER Scores

The WER metric was used for the edit distance. It is shown in Equation 2, where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and C is the number of correct words.

$$WER = \frac{S + D + I}{S + D + C} \tag{2}$$

Optionally, S, D and I could be weighted at discretion, to suit the particular domain. In (Hunt, 1990), for example, a 0.5 coefficient was proposed for D and I. However, in this case, the unitary weights were not altered.

The LossROUGE model scored the best (lowest) mean HER (29.69 ± 21.85), while the AccBLEU was the worst performer (35.55 ± 46.2). However, a Friedman test showed that there were no significant differences between all the 5 groups (n=80, 20 scores x 4 annotators) of HER scores ($\chi^2(5)=3.05$, p–value=0.55). Detailed HER scores can be seen in Table 1.

For certain input scores all the models produced an output that consistently required more edits, which can be seen in Figure 2. The "challenging" inputs

[3] These were: Loss, Token Accuracy, Bilingual Evaluation Understudy (BLEU) (Papineni, Roukos, Ward, & Zhu, 2002), Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), Perplexity (Brown, Pietra, Mercer, Pietra, & Lai, 1992), and Word Error Rate (WER) (Klakow & Peters, 2002).

[4] toloka.yandex.com

mostly differed in their rhythm (they contained more dotted notes and more types of durations) than the "easy" inputs. For example, AccBLEU's mini-score no.1747 gets consistently fewer edits, featuring a selection of proper counterpoint species and consistent tertial harmony on strong metrical positions. Conversely, in mini-score no.2577, shown in Figure 3, this model's hypothesis features broken septuplets (whereas all other models responded with triplets) and poor harmony (clusters of major and minor seconds), which led to an unusual number of edits in this case.

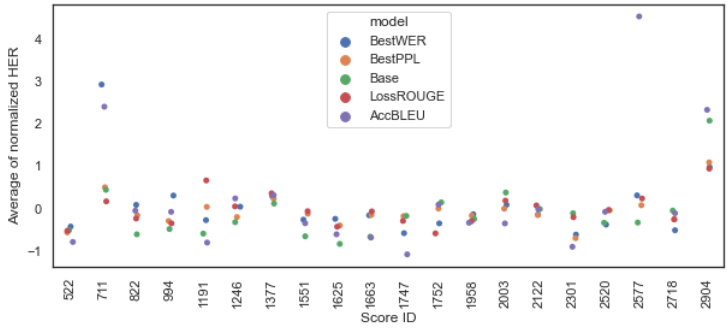

Fig. 2: Averaged across annotators, normalised by annotator HER, grouped by input (musical) score.

After calculating the HER score for each measure of each mini-score, and averaging over all models and all annotators, we noted that for most mini-scores (16 out of 20) the last bar of the hypothesis was the most edited.

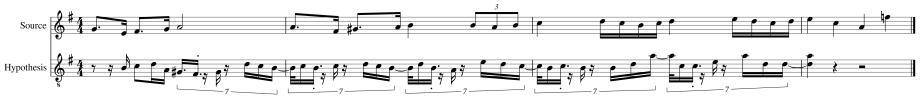

Fig. 3: The most edited hypothesis of the batch evaluated by the annotators.

3.3 Inter-Annotator Agreement

Inter-annotator agreement was calculated using both the *Krippendorff's alpha coefficient* (Krippendorff, 2004) and the *intraclass correlation coefficient* (ICC) for a fixed set of annotators rating each target (Bartko, 1966). All models apart from the LossROUGE (which had the lowest agreement) ranged between poor to moderate agreement. Krippendorff's alpha and ICC for the overall inter-model agreement were 0.388 and 0.411, respectively. When calculated on the HER scores normalized per annotator, the values stood at 0.483 and 0.61.

3.4 Insights

Table 1 summarises the results of this small study. These can be expressed as follows: based on inter-annotator reliability, there was moderate agreement in deeming the model optimised for the perplexity metric the most successful in producing a valid contrapuntal part to a musical query. The base model achieved comparable agreement and came second in HER scores. There was also similar agreement in judging the model optimising Token Accuracy and BLEU as the worst performing of all. Finally, the model optimised for Loss and ROUGE metrics was the least agreed upon, which invalidates its best mean HER score.

Model	**HER** (Mean±Std)	**Krippendorff's** 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 1: Intra-model HER scores and inter-annotator agreement.

Upon closer inspection at a bar level, we observed heavier editing in the final measure of the mini-scores, which could suggest that annotators considered the mini-scores as independent compositions (as opposed to a part of a larger section) which required stronger cadential movement and/or harmonic closure.

4 Baseline

MGEVAL was used as a baseline. The exhaustive cross-validation based on intra and inter-test measurements somewhat confirmed the automatic computational linguistics metrics results (see Section 3), in that there weren't significant differences between the models. As for Kullback–Leibler Divergence (KLD) and Overlap Area (OA), both Base and AccBLEU seemed to do consistently better than the rest, with BestPPL featuring in a couple of metrics, too. There are similarities between these observations and the inter-annotator agreement scores, although discordant with respect to AccBLEU. It is difficult to compare these results to HER's, since MGEVAL is based on the notion of similarity between the training corpus and the generated set. In HER, this concept is only relevant insofar as the annotators' specific domain knowledge is concerned. Thus, Yang & Lerch's approach is clearly more apt to evaluate a set of generations with respect to a training corpus, whereas HER is particularly suitable for the evaluation of single generations. It is worth mentioning that HER scores were computed over a handful of test scores, whereas MGEVAL was computed over the whole test set. Arguably, feature-based methods could be employed alongside HER, rather than in a mutually exclusive fashion, since they contribute different strengths and insights. Some examples of this baseline evaluation are shown in Figure 4 and Table 2. It should be noted that features such as *PC/bar*, *NC/bar*

and *PCH/bar* require a fixed number of bars, which prevented us from using MGEVAL in its complete form.

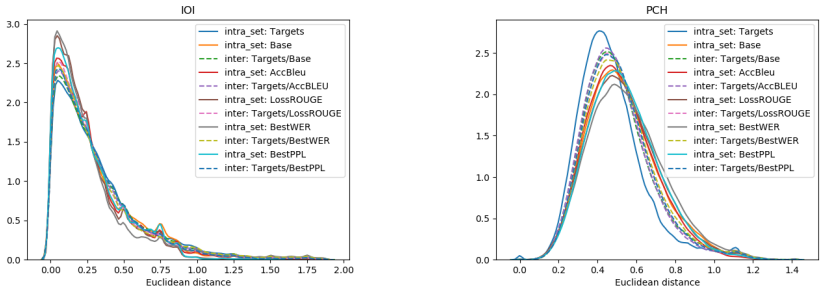

Fig. 4: Intra and inter-test measurements on average inter-onset interval (IOI, left) and pitch class histogram (PCH, right).

	NC		NLH		NLTM		PCH		PC		PCTM		PR	
	KL	OA	KL	OA	KL	OA	KL	OA	KL	OA	KL	OA	KL	OA
Base	.0026	.9380	**.0292**	**.8977**	.0266	**.9696**	**.0144**	.9293	.7907	.8401	.1857	.9652	.0318	**.9548**
AccBLEU	**.0008**	**.9549**	.0346	.8922	.0107	.9540	.0148	**.9314**	**.7730**	**.8758**	**.0050**	**.9792**	.0060	.9329
LossROUGE	.0011	.9458	.0588	.8573	.0166	.9498	.0168	.9255	.8333	.8099	.0075	.9724	.0458	.9380
BestWER	.0011	.9499	.0511	.8665	.0162	.9517	.0237	.9101	.7810	.8204	.0069	.9735	.0056	.9348
BestPPL	.0017	.9430	.0510	.8674	**.0104**	.9576	.0174	.9234	.8474	.8461	.0079	.9727	**.0049**	.9355

Table 2: Kullback–Leibler Divergence (KLD) and Overlapping Area (OA) between the models' dataset intra-set PDF and the inter-set PDF. Shown for notes used (NC), note length histogram (NLH), note length transition matrix (NLTM), pitch class histogram (PCH), used pitch (PC), pitch class transition matrix (PCTM), and pitch range (PR).

5 Conclusion

We described a novel procedure for the evaluation of generative music systems with symbolic output. Such a method is not devoid of subjective judgement, since no annotator is likely to make exactly the same edits as another, however, subjectivity is measured by inter-annotator reliability tests, giving a possibility to remove unreliable annotators. These tests provide an interpretation map for the reading of HER scores, whereby one can verify if averages and variances are sensible. If desired, HER scores can be obtained at a bar level, providing deeper insights into which behaviours of the system are particularly problematic or, conversely, musical. We contend that the HER procedure could be applied to a wide range of generative (symbolic) music systems allowing, conditioned upon re-scaling the HER scores to a common range, comparison between them. Given the data used in this study, the referenced-targets and the edit distance scores were obtained using mostly monophonic parts; thus, the feasibility of HER on more complex textures, while theoretically possible, awaits proof in practice.

References

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

Artstein, R. (2017). Inter-annotator Agreement. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 297–313). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-024-0881-2_11

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

Becker, J. (1986). Is Western Art Music Superior? *The Musical Quarterly*, *72*(3), 341–359.

Brown, P. F., Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D., & Lai, J. C. (1992). An Estimate of an Upper Bound for the Entropy of English. *Comput. Linguist.*, *18*(1), 31–40.

Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. *Journal of New Music Research*, *45*(3), 239-251. (PMID: 28190932) doi: 10.1080/09298215.2016.1200631

Gjerdingen, R. O., & Perrott, D. (2008). Scanning the Dial: The Rapid Recognition of Music Genres. *Journal of New Music Research*, *37*(2), 93-100. doi: 10.1080/09298210802479268

Hunt, M. J. (1990). Figures of merit for assessing connected-word recognisers. *Speech Communication*, *9*(4), 329 - 336. doi: https://doi.org/10.1016/0167-6393(90)90008-W

Kalonaris, S., McLachlan, T., & Aljanaki, A. (2020). Computational Linguistics Metrics for the Evaluation of Two-Part Counterpoint Generated with Neural Machine Translation. In *Proceedings of the First Workshop on NLP for Music and Audio.* International Society for Music Information Retrieval Conference (ISMIR).

Klakow, D., & Peters, J. (2002). Testing the Correlation of Word Error Rate and Perplexity. *Speech Commun.*, *38*(1), 19–28. doi: 10.1016/S0167-6393(01)00041-3

Koops, H. V., de Haas, W. B., Burgoyne, J. A., Bransen, J., Kent-Muller, A., & Volk, A. (2019). Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, *48*(3), 232-252. doi: 10.1080/09298215.2019.1613436

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology.* Sage.

Lin, C.-Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computer and the Humanities*, *24*, 161–175. doi: https://doi.org/10.1007/BF00117340

Navarro, G. (2001). A Guided Tour to Approximate String Matching. *ACM Comput. Surv.*, *33*(1), 31–88. doi: 10.1145/375360.375365

Nichols, E. P., Kalonaris, S., Micchi, G., & Aljanaki, A. (2021). Modeling Baroque Two-Part Counterpoint with Neural Machine Translation. In *Proceedings of the International Computer Music Conference.* Santiago, Chile: International Computer Music Association. (Rescheduled from 2020 due to the covid-19 pandemic. Preprint available: `https://arxiv.org/abs/2006.14221`)

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th annual meeting of the association for computational linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi: 10.3115/1073083.1073135

Ritchie, G. (2019). The Evaluation of Creative Systems. In T. Veale & F. A. Cardoso (Eds.), *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems* (pp. 159–194). Cham: Springer International Publishing. doi: 10.1007/978-3-319-43610-4_8

Selway, A., Koops, H. V., Volk, A., Bretherton, D., Gibbins, N., & Polfreman, R. (2020). Explaining harmonic inter-annotator disagreement using Hugo Riemann's theory of 'harmonic function'. *Journal of New Music Research*, *49*(2), 136-150. doi: 10.1080/09298215.2020.1716811

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Weischedel, R. (2006). A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Translation in the Americas (AMTA).*

Sturm, B., & Ben-Tal, O. (2017). Taking the Models back to Music Practice: Evaluating Generative Transcription Models built using Deep Learning. *Journal of Creative Music Systems*, *2*(1). doi: https://doi.org/10.5920/JCMS.2017.09

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 5998–6008). Curran Associates, Inc.

Yang, L.-C., & Lerch, A. (2018). On the evaluation of generative models in music. *Neural Computing and Applications*, *32*(9), 4773–4784. doi: https://doi.org/10.1007/s00521-018-3849-7