A Unified Evaluation of Expressive Generative Models and Steerable Interfaces for Music Creation[1]

Anonymous Author(s)

ABSTRACT

ML models are becoming more expressive and capable of generating music with long-range coherence. At the same time, better HCI interfaces for controlling them can promote feelings of ownership. While these parallel efforts are aimed at empowering the end-user, less is known about how both ML models and HCI interfaces can impact a creator's subjective experience, and how people objectively perform on a creative task, such as composing music to express an emotion. In this study, we jointly evaluate the impact of ML and HCI advances in empowering co-creation through a common task and measure of expressing and communicating emotion in music. Our study is distinguished in that it measures communication through both composer's self-reported experiences, and how listeners evaluate this communication through the music. In an evaluation study with 26 composers creating 100+ pieces of music and listeners providing 1000+ head-to-head comparisons, we find that more expressive models and more steerable interfaces are important and complementary ways to make a difference in composers communicating through music and supporting their creative empowerment.

KEYWORDS

quantitative methods; generative models; steering interfaces; human-ai co-creation;

ACM Reference Format:

Anonymous Author(s). 2022. A Unified Evaluation of Expressive Generative Models and Steerable Interfaces for Music Creation[1]. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

There is an increasing interest from machine learning (ML) and human computer interaction (HCI) communities in empowering creators with better generative models and more intuitive interfaces with which to control them. In the domain of music, ML researchers have focused on training models capable of generating pieces with increasing long-range structure and musical coherence [4, 8], while HCI researchers have separately focused on designing better steering interfaces that support user control and ownership through overcoming AI-induced information overload and non-deterministic model outputs [5].

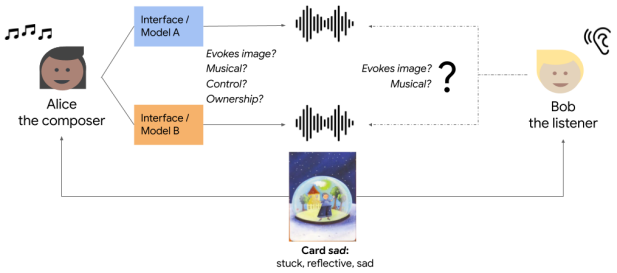

Figure 1: Our study method uses Expressive Communication as a unifying task and measure. For the creative task, composers use different versions of a generative AI tool to create music with the goal of communicating a particular image and human emotion. It supports objective measurement by using an outside listener to judge how the created music better evokes the intended imagery and emotions.

While the ML and HCI communities have similar aspirations for generative modeling, interdisciplinary collaborations have been limited by mismatches in the way each field evaluates progress. One the one hand, many ML researchers desire their models to be useful for creators, but models are not directly measured on metrics *downstream* from training, such as empowering individuals to achieve their creative goals. Instead, they are measured using proxy metrics that are easier to automate, such as the ability to generate realistic samples that imitate the training data. On the other hand, HCI researchers will conduct user studies with creators to evaluate whether an interactive generative tool is more controllable and promotes feelings of collaboration and ownership. While these studies focus on the experience of the creator as measured through self-report [5, 6, 12], an equally important metric is how the products of creation can have an effect on an outside audience. Within the domain music, it remains untested whether better steering interfaces for generative models can empower users of these tools to create compositions that better evoke an emotion or sound more musical, as judged by listeners.

This led us to ask the question: *"How do recent ML and HCI advances in generative tools impact downstream creative tasks, as measured by composer's subjective self reports and by objective judgements by outside listeners?"*. Our study uses the idea of Expressive Communication as the downstream task and metric, as illustrated in Figure 1, where a novice composer uses generative tools to create pieces of music that express a particular feeling and image, while an outside listener judges which music actually best evokes that feeling and image. We compare between two generative models capable of different degrees of long-range structure and musical

[1]This workshop paper is a shortened summary of the full IUI'22 paper []
[2]This work was completed during the first author's summer internship at Google.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

coherence, and also between two different interfaces capable of different degrees of steering and iterative composition.

The composers in our study created 100+ musical phrases [3], expressing the imagery and words in the set of illustrations in Figure 3, which were then evaluated by listeners in 1000+ head-to-head comparisons. Our results show that both the ML and HCI approaches (developing better pretrained models and better steering interfaces, respectively) are important and complementary ways to support composers in both communicating through music and feeling empowered in the process of co-creating with generative models. Our results also shed light on how biases in a pretrained model capabilities such as stronger coherence can make feelings such as fear more difficult to express with curation of random samples alone, and how the addition of steering interfaces can help to overcome model biases by creating samples that are less likely from the model, but more aligned with the user's expression and musical goals.

2 COMPARISON OF GENERATIVE MODELS AND STEERABLE INTERFACES FOR MUSIC CREATION

In this unified evaluation, we setup two experimental comparisons: an ML model comparison and an HCI steering interface comparison.

2.1 Comparing Expressiveness of Models

In the model comparison, users made the music that expressed an emotion (e.g., sad) using two different generative tools. While both generative tools had the same interface for curating music, they used different models. For our studies, we compared two trained autoregressive generative models capable of generating polyphonic piano music:

- **Performance RNN** [7] is the baseline and less expressive model. It is trained on a smaller and more dramatic piano performance dataset (MAESTRO) [2], and has less capacity in the model architecture to model the distribution of music.
- **Music Transformer** [4] is the more expressive model capable of building upon and developing expressive themes and motifs through long-range coherence. It is trained on a larger and more melodic piano performance dataset (YouTube) [10], and has larger capacity to model the distribution of music.

2.2 Comparing Steerability of Interfaces

In the interface comparison, users made music that expressed a different emotion (e.g., conflict) using two different generative tools. While both generative tools used the same generative model, they provided different interfaces for control.

The design of the two interfaces that provide different levels of steering is illustrated in Figure 2. We implemented a baseline and less steerable interface called the **Radio Interface**, inspired by a common approach when interacting with an "end-to-end" model of generating large quantities of samples and curating to find the best results [3]. A composer using the Radio Interface requests randomly generated options of the full-length phrase of music, and selects

[3]Listen to the music participants composed using generative tools for different imagery and words here: https://storage.googleapis.com/expressive-communication/index.html

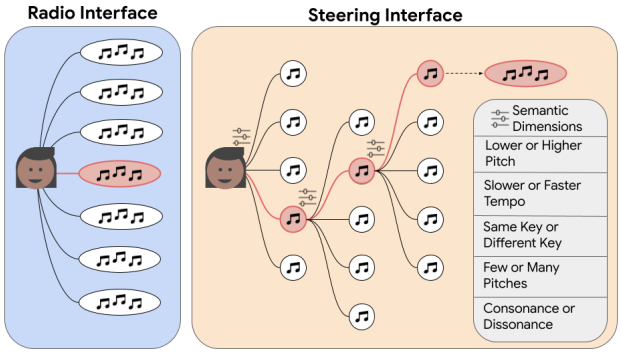

Figure 2: We compared two interfaces for creating music with an autoregressive generative model: the Radio Interface and the Steering Interface.

Figure 3: We selected 5 image cards from the board game Dixit [11]. To ensure that composers and listeners had a common interpretation of the card, we attach three keywords to each card.

the one that best matches her goals. We implemented a **Steering Interface** which embodied two interaction principles for creating more steerable generative tools, as studied in [5]: generating music chunk-by-chunk; and having controls to constrain the generation along semantically-meaningful dimensions. A composer using the Steering Interface will explore and select three total generated chunks for the start, middle, and end of their musical phrase For exploring any chunk, they can request a randomly generated set, or they can choose to constrain the generation along semantic dimensions (e.g., lower or higher pitch; consonance or dissonance). After selecting any chunk, subsequent chunks are generated as a continuation of the previous chunks, in an autoregressive fashion.

3 EVALUATION STUDY

In this study, we ask *"How does recent ML progress in more expressive models and HCI advances in more steerable interfaces impact **creative empowerment** and **effectiveness in communicating emotions**, as measured by composer's subjective self reports and objective judgements by outside listeners?"*.

3.1 Composer Study Method

3.1.1 Study Setup. The 26 participants who completed the study included 12 females and 14 males, ages 24 - 60 (μ = 37). The composer sessions were conducted remotely using video call and screensharing. Each user was given an overview of the goals of the study

and the expressive communication game (10 minutes). Participants were assigned to a comparison (interface comparison or model comparison) according to the counterbalanced ordering and completed a guided tutorial of the system (15 minutes). In the first comparison, participants were assigned a card and were asked to compose music that reflected the imagery and words of the card using each of the systems being compared for the comparison (10 minutes per system). Users were observed while composing using a think-aloud procedure. Finally, they answered a post-comparison questionnaire and completed a semi-structured interview comparing their experiences (10 minutes). This procedure was repeated for the second experimental comparison (40 minutes). At the end, they answered several questions about their overall experience composing in this expressive communication game (10 minutes).

3.1.2 Measures and Analysis. To answer our RQs about the impact of different generative tools on composers' experiences, we used a combination of quantitative survey ratings, and qualitative data from interviews and talk-alouds. For our quantitative measures, we asked participants to rate six survey measures for both system versions they had used. All measures below were rated on a 7-point Likert scale (1=Strongly Disagree, 7=Strongly Agree).

We asked three questions about the perception of the final piece of music. **Expression**: Users rated *"I feel confident that my composition created with System X expresses the imagery/ideas of the card I chose."* **Communication**: Users rated *"I feel confident that others will be able to rank the Card I chose as high after listening to the composition created with System X."* **Musical Coherence**: Users rated *"I feel the composition I created with System X feels musically coherent."*

We asked three questions about participants' attitudes towards the generative tool and their composition experience using the tool. **Ownership**: Users rated *"Using System X, I felt the composition created was mine."* **Control**: Users rated *"I felt I had control creating the composition when using System X."* **Efficacy**: Users rated one item from the Generalized Self-Efficacy scale [9] rephrased for music composition, which asked *"Using System X, I could find several solutions to achieve my goals for the composition."*

To analyze these quantitative differences, we conducted a two-sample paired t-test, with the null hypothesis that the mean of their differences is zero. When testing for significance, we used a Bonferroni-corrected threshold of 0.00416 to correct for the 12 tests performed. For our qualitative analysis, we coded the quotes from interviews and talk-aloud sessions using a deductive, thematic analysis [1] according to the various dimensions we asked in quantitative results (e.g., expression; musical coherence; control), and grouped by the type of model or interface used.

3.2 Listener Study Method

3.2.1 Study Setup. We recruited 20 unique listeners for the listener study from an online crowd work platform. Each listener made head-to-head comparisons for all 51 pairs of musical samples created by the participants in our composer study (26 interface comparisons and 25 model comparisons). In total, our listeners provided 1020 head-to-head ratings comparing the pairs of music compositions.

We asked listeners to compare the two phrases of music which were created by the same composer, for the same card. We chose this to control for the variations in composers interpretations of a given card, and to control for variations in how easy it might be to express one card vs. the other. The ordering of the music options were randomized to prevent an association with ordering and experimental conditions.

3.2.2 Measures and Analysis. After being given two musical samples to compare, listeners rated two questions that asked *"Which one of these musical excerpts **most evokes the feelings** of the words and imagery on the card?"* and *"Which one **sounded more musical**"* and answered on a 5 point balanced scale (*"Strong preference for option 1"*, *"Weak preference for option 1"*, *"No preference"*, *"Weak preference for option2"*, and *"Strong preference for option2"*). In preparation for conducting our analysis, we converted this scale to a numerical scale from $[-2, 2]$. For for the model comparison, positive values correspond to a preference for Music Transformer; for the interface comparison, positive values correspond to a preference for the Steering Interface. To analyze these listener ratings, we used a one-sample t-test, with the null hypothesis that there would be no preference ($\mu = 0$) for either of the models or interfaces.

3.3 Composer Results

3.3.1 Model Comparison. The composer self-report ratings for the model comparison is shown in Figure **??**. In our Model Comparison, we found that Music Transformer, the more expressive model, improved novice composers' sense of **ownership** and **efficacy**. Novice composers felt that the generated music was more **musically coherent**. Expressive Models helped composers efficacy in finding multiple options that could achieve their creative goals. Several said this was because samples communicated a *"single thread"* which *"repeats and builds upon a small theme,"* ultimately helping to evoke a clear image and mood (P5). In contrast, music made with PerformanceRNN often had interjections, like abrupt changes in tempo and tonality, making it hard to evoke specific emotions.

3.3.2 Interface Comparison. The composer self-report ratings for the interface comparison is shown in Figure **??**. The more Steerable Interface made composers feel empowered, by improving their sense of **ownership**, **control**, and **efficacy**. The Steering Interface allowed some participants to thinking about direction of the music in terms of a narrative, with different elements in different chunks. For example, P7 said they *"could more piece it together"* through the ability to make choices for each chunk, and said they could *"compose a flow, a narrative"* and *"think about the direction it takes"*. It allowed users to provide more input to the generation, as compared to the Radio Interface where people were just choosing different generated examples. As one participant using the baseline Radio Interface said, *"since I showed my preference only a little bit in choosing from the different examples, I don't feel it was mine"* (P4).

3.4 Listener Results

3.4.1 Model Comparison. Listeners felt compositions made with **Music Transformer** better evoked the feelings of the card over compositions made with **Performance RNN**, where the difference was statistically significant ($\mu = .24$, $t = 3.318$, $p < 0.001$). In addition, listeners rated compositions made with Music Transformer

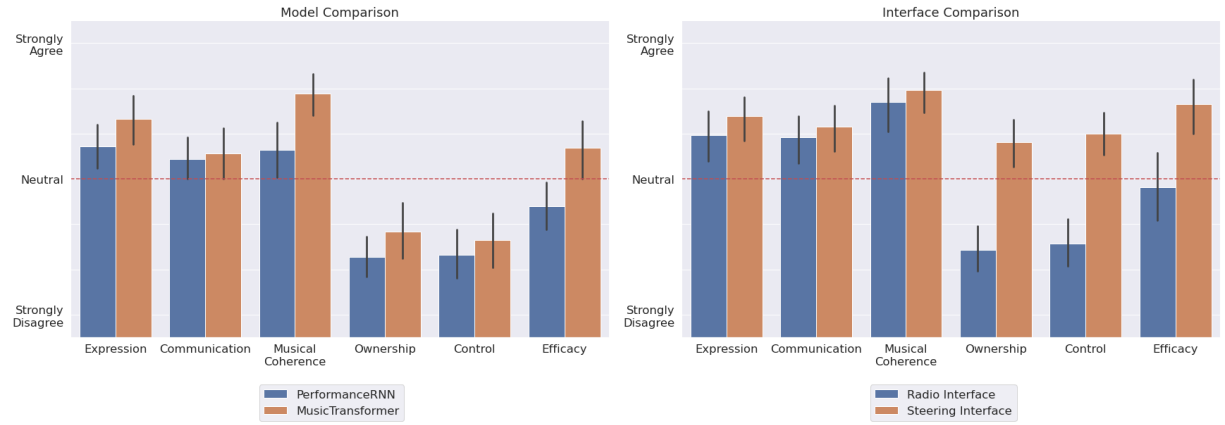

Figure 4: Composers answered questions about on the final compositions and about their experiences with regards to (a) the different models, and (b) the different interfaces.

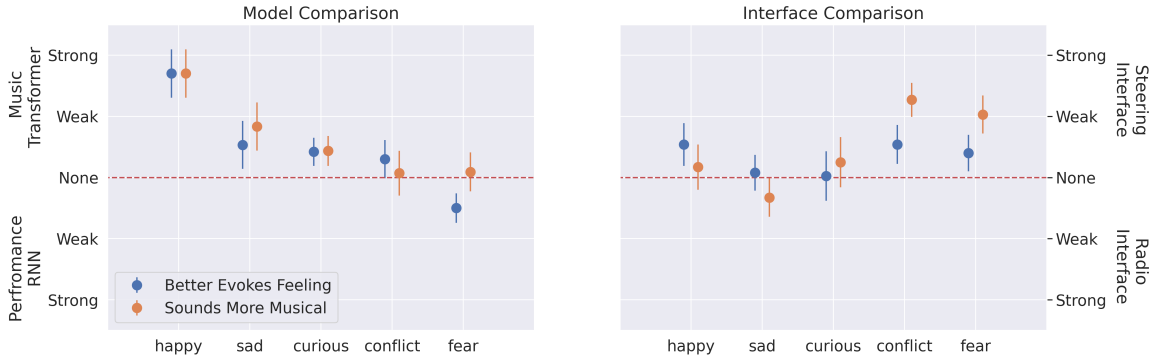

Figure 5: Listener ratings, comparing which model and interface best evoke the target feeling and/or sound more musical.

to sound more musical than those made with Performance RNN ($\mu = 0.378$, $t = 4.74$, $p < 0.00001$).

3.4.2 Interface Comparison. Listeners on average felt compositions made with the **Steering Interface** better evoked the feelings of the card over compositions made with the **Radio Interface**, where the difference was statistically significant ($\mu = 0.31$, $t = 4.09$, $p < 0.0001$). Additionally, listeners felt compositions created with the **Steering Interface** sounded more musical than those made with the **Radio Interface** ($\mu = -0.5846$, $t = 3.472$, $p < 0.001$).

3.5 Steering Interfaces correct Model Biases

3.5.1 Biases in Pretrained Models Help to Better Evoke Some Feelings Over Others. Comparing model preferences by card (Figure 5, left) exposes a bias in pretrained models, where curated random samples from Music Transformer evoke the feelings of happy, sad, curious, and conflict, while Performance RNN samples better evoke fear. Interestingly, effectively evoking the feeling is strongly correlated with sounding more musical, revealing that pretrained Music Transformer has a bias towards coherent musical output which is more aligned with straightforward feeling such as happy or sad. Since fear can often be expressed in sudden changes, or surprises, the model that expresses coherent ideas is less likely to output music that matches these qualities.

3.5.2 Steering Interfaces are More Helpful When Expressing Feelings that are Misaligned with Model Biases. In the second graph (Figure 5, right), we see that adding steering interfaces to the Music Transformer model helps the most for the emotions that it was originally not suited for, such as conflict and fear. That is because steering helps with finding samples not as likely by random curation alone, but that does express the desired emotion. As an example of a steering strategy to express the feelings of fear, a composer selected a first chunk that is *"slower and has less energy to signal something impending that hasn't happened yet"*, added more feelings of ominous by setting the semantic parameters to *"slower, lower, and different key"*, and steered the third chunk to be faster and much lower in the end to *"build tension at the end to signal that the character is being attacked by the killer plants"* (P3).

REFERENCES

[1] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[2] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2018. Enabling factorized piano music modeling and generation with the MAESTRO dataset. *arXiv preprint arXiv:1810.12247* (2018).

[3] Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J Cai. 2020. AI song contest: Human-AI co-creation in songwriting. *ISMIR* (2020).

[4] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer. *ICLR* (2018).

[5] Ryan Louie, Andy Coenen, Cheng-Zhi Anna Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3313831.3376739

[6] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 649, 13 pages. https://doi.org/10.1145/3173574.3174223

[7] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. 2020. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications* 32, 4 (2020), 955–967.

[8] Christine Payne. 2019. MuseNet. https://openai.com/blog/musenet. Accessed: 2020-05-04.

[9] Ralf Schwarzer and Matthias Jerusalem. 1995. Generalized Self-efficacy Scale. *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs* 1, 1 (1995), 35–37.

[10] Ian Simon, Cheng-Zhi Anna Huang, Jesse Engel, Curtis Hawthorne, and Monica Dinculescu. 2019. Generating Piano Music with Transformer. *Magenta Blog: https://magenta.tensorflow.org/piano-transformer* (2019).

[11] Wikipedia contributors. 2019. Dixit (card game) — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Dixit_(card_game)&oldid=908027531. [Online; accessed 19-September-2019].

[12] Yijun Zhou, Yuki Koyama, Masataka Goto, and Takeo Igarashi. 2021. Interactive Exploration-Exploitation Balancing for Generative Melody Composition. In *26th International Conference on Intelligent User Interfaces*. 43–47.