IteraTTA: AN INTERFACE FOR EXPLORING BOTH TEXT PROMPTS AND AUDIO PRIORS IN GENERATING MUSIC WITH TEXT-TO-AUDIO MODELS

Hiromu Yakura
University of Tsukuba
Tsukuba, Japan
hiromu.yakura@aist.go.jp

Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
m.goto@aist.go.jp

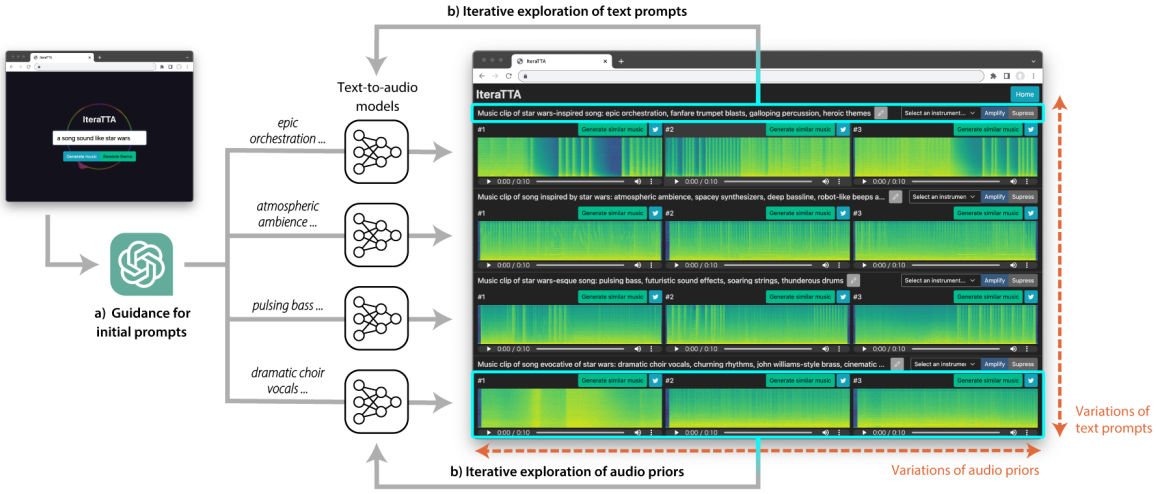

Figure 1. IteraTTA is an interface dedicated for allowing novice users to show their creativity in text-to-audio music generation processes. It provides a) computational guidance for constructing initial prompts and b) dual-sided iterative exploration of text prompts and audio priors.

ABSTRACT

Recent text-to-audio generation techniques have the potential to allow novice users to freely generate music audio. Even if they do not have musical knowledge, such as about chord progressions and instruments, users can try various text prompts to generate audio. However, compared to the image domain, gaining a clear understanding of the space of possible music audios is difficult because users cannot listen to the variations of the generated audios simultaneously. We therefore facilitate users in exploring not only text prompts but also audio priors that constrain the text-to-audio music generation process. This dual-sided exploration enables users to discern the impact of different text prompts and audio priors on the generation results through iterative comparison of them. Our developed interface, IteraTTA, is specifically designed to aid users in refining text prompts and selecting favorable audio priors from the generated audios. With this, users can progressively reach their loosely-specified goals while understanding and exploring the space of possible results. Our implementation and discussions highlight design considerations that are specifically required for text-to-audio models and how interaction techniques can contribute to their effectiveness.

1. INTRODUCTION

Recent advances in generative machine learning techniques open up novel ways for a diverse group of individuals to engage in creative processes [1, 2]. Specifically, music generation models can foster creative expression among novice users, who may not necessarily possess formal musical knowledge [3, 4]. Consequently, several approaches have been proposed to enable users to control various musical attributes of generated audios, such as specifying the note or rhythm density [5, 6] and chord progression [7–9]. Text-to-audio models [10, 11] are promising in terms of allowing users who are not familiar with the concepts of such musical attributes to generate their own sounds.

Nevertheless, there are still several gaps toward deploying such models to support the creativity of novice users. For example, the models rely on annotated labels of music clips presented in their training datasets [12–14], which primarily consist of musical descriptions such as genres, instruments, and moods. Therefore, providing such infor-

mation as a text prompt is crucial for enabling fine-grained control over generated music audios. However, this may prove challenging for novice users due to disparities in artistic vocabulary among individuals with varying levels of musical knowledge [15]. Experimentally, it has been suggested that non-musicians tend to rely more on abstract concepts, such as the pleasantness or complexity of music, when appreciating musical pieces [16], which may pose difficulties in fully exploring various text prompts.

Moreover, understanding the space of possible results is also challenging, particularly when compared to the use of text-to-image models. In text-to-image generation, users can look over various generation results at a glance, which fosters their understanding of the space and helps them decide on directions to explore [17]. From the perspective of explainable AI (XAI), we can say that such results serve as *explanations by example* [18] because the results implicitly invite the users to infer the behavior of the models. However, in text-to-audio generation, users cannot simultaneously listen to multiple generation results, thus impeding their comprehension and ability to efficiently explore the space. These points imply that specific design considerations are necessary to fully leverage the potential of text-to-audio models and exploring them would also provide a new perspective in terms of XAI.

In this paper, we introduce IteraTTA, an interface dedicated to the text-to-audio (TTA) music generation processes of novice users. This interface enables iterative exploration of both text prompts and audio priors, allowing users to gain a comprehensive understanding of the space of possible results by sufficiently constraining the generation processes. We constructed this interface based on our observations and related literature on creativity support, which emphasize the importance of 1) computational guidance for constructing initial prompts and 2) dual-sided iterative exploration of text prompts and audio priors. Moreover, we deployed the interface as a publicly-available Web service and analyzed the diverse ways in which users utilized it in their creative processes. Our results and discussions shed light on ways to utilize models developed in the MIR community to unleash the creativity not only of expert users [19] but also of individuals with varying degrees of musical knowledge.

2. RELATED WORK

2.1 Music Generation Techniques

Music generation has been one of the central topics with the MIR community [20–24], and recently, generative machine learning techniques have been widely employed for this purpose [24, 25]. While methods for symbolic music generation that output MIDI files have been popular [26–32], some methods use generative models to directly output audio, leveraging their expressiveness [33–36]. For example, Jukebox [33] and RAVE [34] use variational autoencoders and autoregressive models trained on large-scale music datasets to generate diverse music audios.

Controllability in music generation has been also emphasized [5–9, 37–39] because it is vital to open up its applications for supporting users' creative processes [40, 41]. For instance, Music FaderNets [5] allows users to modify the rhythm and note densities of generation results, while Music SketchNet [6] enables them to specify pitch contours and rhythm patterns. Wang *et al.* [7] and Dai *et al.* [8] have proposed methods to further constrain the chord progression of generation results. However, as mentioned in Section 1, users are not always familiar with such concepts, and then, they would have difficulties in using these methods to output music audios they want to generate. We acknowledge that some methods [38, 39] provide perceptual control that does not require extensive musical knowledge: emotion-based musical generation. Nevertheless, they are based on Russell's valence-arousal model [42] consisting of four classes, which limits the range of controls and may hamper users' agency [43] when the methods are used to support their creative processes.

In this context, recent text-to-audio models [10, 11] can be an effective tool for such novice users. These models learn the relationship between music audios and their text descriptions (more specifically, latent representations encoded from the descriptions by RoBERTa [44]) and use it to guide results in generating new audios from an inputted text (*i.e.*, text prompt). As RoBERTa can encode text prompts with variable length and content, the models can provide flexible control without requiring specific musical knowledge of rhythm patterns or chord progressions. Moreover, they allow users to constrain generation results not only by text prompts but also by audio priors, ensuring that the results have similar characteristics to the priors. For example, the diffusion model [45] employed by AudioLDM [11] usually uses Gaussian noise for the seed of its generation process, but by using a noise-infused audio prior, we can obtain generation results preserving the characteristics of the provided audio.

Here, text-to-image models that use similar schemes have been shown to unleash the creativity of novice users, allowing them to iteratively explore open-ended variations of text prompts [17] and customize their intermediate results by specifying image prior constraints [46]. Similarly, text-to-audio models can be leveraged to provide users with such iterative exploration or customization. However, we also expect that text-to-audio music generation processes may pose several specific difficulties, as explained in Section 1. Therefore, we explored how interaction techniques can address these challenges by developing an interface dedicated to text-to-audio models.

2.2 Interfaces for Music Generation

There is a series of research on building interfaces to let users interact with music generation techniques effectively [47–53]. MySong [47], for instance, involves a music accompaniment generation model, with which users can control the happiness or jazziness of generation results. Louie *et al.* [49] proposed an interactive interface for novice users so that they can use a symbolic music generation technique with control of happiness or randomness.

The interface also allows users to constrain generation results by providing music priors, which was experimentally confirmed to be effective in iteratively refining the results. Zhou *et al.* [52,53] utilized a user-in-the-loop Bayesian optimization technique to enable novice users to iteratively explore melodies composed by a generative model.

These interfaces underscore the significance of providing controls and supporting iterative exploration in facilitating the creativity of novice users using music generation techniques. Consequently, the provision of recent text-to-audio models to novice users would be highly suitable for this purpose, as they offer more flexible control, compared to using several parameters such as happiness, while also allowing the use of audio priors. Our paper contributes to this series of research by examining design considerations of interfaces for text-to-audio music generation processes, aiming to expand the scope of applications of recent techniques developed in the MIR community.

3. DESIGN REQUIREMENTS

As stated in Section 1, our goal is to leverage text-to-audio models to facilitate the creative expression of novice users regardless of their musical knowledge. To this aim, we embarked upon an examination of potential challenges that these users may encounter during text-to-audio music generation processes and subsequently derived a set of design requirements to address these issues. Guided by the principles of human-computer interaction, we utilized the think-aloud protocol [54, 55] by involving three volunteers who self-reported that they possessed no formal musical training beyond compulsory education. Specifically, we provided the volunteers with access to one of the latest text-to-audio models [11] on Google Colab using its official implementation [1] , which enabled them to provide any text prompts and subsequently listen to three music audios generated from the text prompts. Here, since the remotely-participated volunteers were Japanese speakers recruited via word-of-mouth communication, we told them that they can use DeepL Translator to translate text prompts into English to obtain better results with the model that is mainly trained on the dataset with English text labels [12–14]. They freely used the model for approximately 30 minutes while sharing their screens on a video call and verbalizing their thoughts and feelings. This allowed us to identify the challenges that they encountered and the factors that contributed to these challenges. We then conducted semi-structured interviews to validate the challenges identified and to gain further insight into the reasons behind them. Their responses were analyzed based on open coding [56], which yielded the following design requirements in line with the existing literature on creativity support.

3.1 Computational guidance for constructing initial prompts

We observed that the volunteers frequently encountered difficulty in formulating appropriate text prompts to initiate their use of the model. For example, one volunteer entered the phrase "a song sounds like star wars," resulting in audio containing a battle cry with a space-like sound effect. This can be attributed to the characteristics of the text labels in the dataset used to train the model [12–14]. Specifically, the labels of music clips consist primarily of musical descriptions such as genres, instruments, and moods, like: "An orchestra plays a happy melody while the strings and wind instruments are being played [14]." Therefore, providing such a description would be essential to ensure that the model trained on the dataset generates music audio as intended. The volunteer was unable to generate music-like audio until he attempted several prompts and finally entered "solemn music starting with a trumpet fanfare."

In the context of creativity support, two underlying factors could explain the aforementioned observation. First, an inherent gap in artistic vocabulary exists between expert and novice users [15]. Without deep musical knowledge, it can be challenging to conceive a precise description of music audios. Additionally, novice users often have loosely-specified goals when starting a creative endeavor [57–59]. They refine their objectives gradually by exploring the space of possible results through iterative exploration [60, 61]. However, the dependency of text-to-audio models on precise descriptions of clearly-defined goals makes it difficult for novice users to initiate such exploration. This suggests that supporting them computationally in constructing initial prompts could potentially facilitate the creativity of novice users.

3.2 Dual-sided iterative exploration of text prompts and audio priors

We also observed that the volunteers encountered challenges in efficiently exploring the generated results. One volunteer who had prior experience with text-to-image models mentioned the point, as:

> *"Unlike text-to-image models, comparing various results at a glance was difficult with the text-to-audio model. So, finding a text prompt reflecting my intention most faithfully became much tough."*

In other words, iteratively trying different text prompts would not necessarily assist users in comprehending the space of potential results, although it is necessary for novice users to refine their loosely-specified goals [60,61]. Therefore, users cannot determine which direction would be closest to their goals and what text prompt to try next. Another volunteer mentioned an issue he faced, as:

> *"I once found a generation result with a good melody, but I wanted to change its tone. So, I added 'with a flute' to its text prompt and regenerated. However, the melody was then completely changed, which was frustrating."*

This implies that we need to let users utilize not only text prompts but also audio priors to constrain the tune of generation results. In sum, supporting the creativity of novice users in text-to-audio music generation processes requires enabling them to efficiently explore variations of both text

[1] https://github.com/haoheliu/AudioLDM

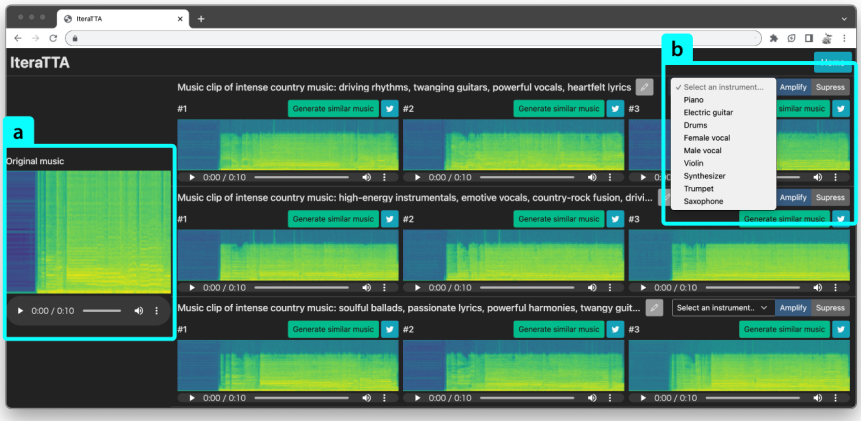

Figure 2. To facilitate the exploration of text prompts and audio priors, IteraTTA allows a) comparison of generation results with an audio prior and b) instant edit of a text prompt.

prompts and audio priors, allowing them to iteratively refine their goals by understanding the space of possible results. This demands us to develop an interface specifically tailored for text-to-audio models to provide such dual-sided exploration of text prompts and audio priors.

4. IteraTTA

Based on the above design requirements, we present IteraTTA, a dedicated interface for text-to-audio music generation processes. It was implemented as a Web-based system, allowing novice users to instantly benefit from the latest text-to-audio models in their creative processes.

4.1 Design

As illustrated in Figure 1, our interface requires users to first input a theme phrase for music audios to generate. The inputted phrase need not include precise musical descriptions since IteraTTA leverages a large language model to derive such descriptions suitable for text-to-audio models using knowledge embedded in the models [62]. Specifically, the interface queries a large language model that "Please give me four variational lists of comma-separated phrases describing what does a music clip of "*[theme phrase]*" sound." It then uses the four responded phrase lists as a variety of the first text prompts to start the music generation processes in parallel. This feature allows novice users to translate loosely-specified goals in their minds into musical descriptions, which can also help them to envisage variations of text prompts to explore.

IteraTTA then generates three music audios for each of the four prompts. The generated audios are arranged in two dimensions (see Figure 1), which enables novice users to understand how different music audios are generated by different text prompts, and also, how different music audios are generated by the same text prompts. This is intended to assist users in identifying which text prompts and audio priors are closely aligned with their goals and which direction is worth exploring. If a user identifies a suitable candidate text prompt, they can customize the prompt and generate new music audios with it. Alternatively, if the user discovers a suitable music audio, they can use it as an audio prior to generate new music audios. In essence, the user can explore the subspace of possible results that are proximate to their goals by constraining either text prompts or audio priors, while gradually refining their goals by themselves.

We have incorporated several features to facilitate the exploration of text prompts and audio priors, as shown in Figure 2. For instance, when a user specifies an audio prior, IteraTTA enables the user to compare generated results with it. It also offers an instant editing feature of text prompts, allowing users to amplify or suppress the sound of a selected instrument. This is achieved by simply adding a phrase of "with strong *[instrument]*" or "with no *[instrument]*" into a text prompt, but it provides an example of how they can modify generation results through prompts.

4.2 Implementation

As mentioned, we developed IteraTTA as a Web-based system to invite novice users for trying music generation with it. For the implementation of its back-end server, we utilized Python with FastAPI and incorporated an API of GPT-3.5 [2] to construct initial prompts, while AudioLDM [11] was employed to generate the music audios. The length of music audios to generate was predetermined at 10 seconds so that our GPU server harnessing an NVIDIA RTX 2080 Ti can afford the generation of 12 audios (3 audios $\times$ 4 prompts) simultaneously. On average, the generation process takes approximately 15 seconds. In addition, we used DeepL API to translate text prompts into English when they were provided in non-English languages because we observed that it led to better results in Section 3. For the front-end interface of IteraTTA, we utilized Vue.js, which enables users to download the generated music audios or share them on Twitter.

[2] We used `gpt-3.5-turbo` of `https://platform.openai.com/docs/models/gpt-3-5`.

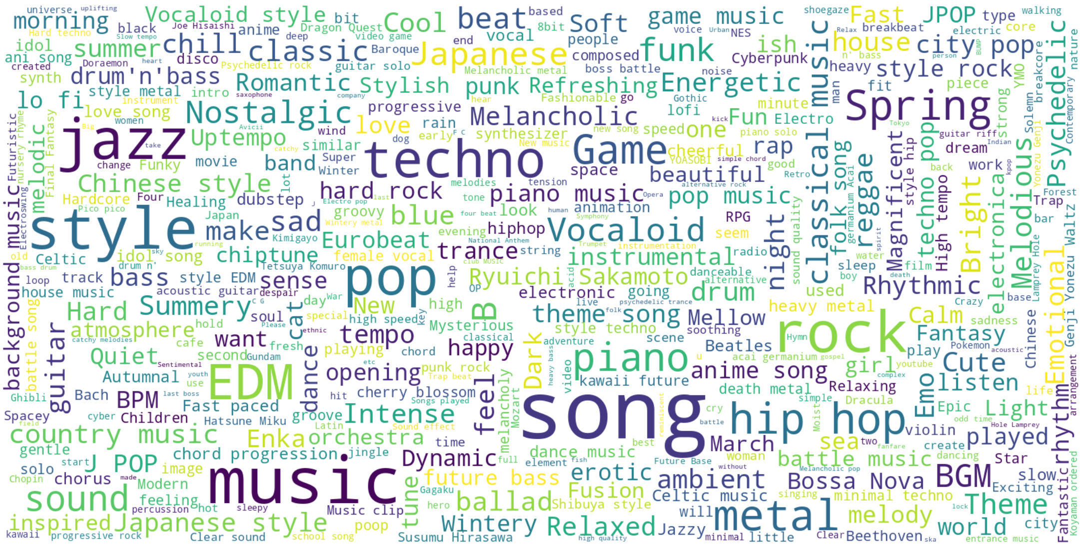

Figure 3. Word cloud of theme phrases the 8,831 users inputted on our Web service.

5. ANALYSIS

To investigate the effectiveness of IteraTTA in supporting diverse users in the wild, we deployed it as a publicly-available Web service in Japanese [3]. Within two days of release, 8,831 users generated 246,423 music audios. In this section, we discuss the insights we extracted from their usage logs and their responses to a form that we put a link to it on the Web service so that they can share their opinion and feedback voluntarily.

5.1 Diversity of theme phrases

We first examined the theme phrases that users inputted to initiate text-to-audio music generation processes and found that they were highly varied. Some users provided music-related phrases, such as "nice city pop" and "cute future bass," while others were more specific, like: "80's hip hop that break dancers would dance to." There were also phrases expressing more abstract ideas, such as "Arabian caves" and "silent dream of a priestess." Figure 3 visualizes the words often used in the translated phrases in the form of a word cloud, showing their diversity.

To explore the role of IteraTTA, we compared the theme phrases inputted by the users and the text prompts derived from them by the large language model to the text labels in the dataset used for training the text-to-audio model. Specifically, we randomly sampled 1,000 cases for each of the theme phrases, text prompts, and text labels [4] and calculated their representation vectors using the same pre-trained model of RoBERTa [44] as the text-to-audio model. We then visualized the distribution of the vectors using t-SNE [63], as presented in Figure 4. This indicates that IteraTTA guided the large language model to derive text prompts that bridged the gap between the diverse users' theme phrases and the text labels in the training dataset. In fact, we found that the large language model successfully derived text prompts containing musical descriptions even from abstract phrases, such as "otherworldly harmonies, delicate strings, minimalistic percussion, dreamlike vocals" for "silent dream of a priestess." These results suggest the effectiveness of guiding the construction of initial prompts to support the creative processes of novice users, as discussed in Section 3.1.

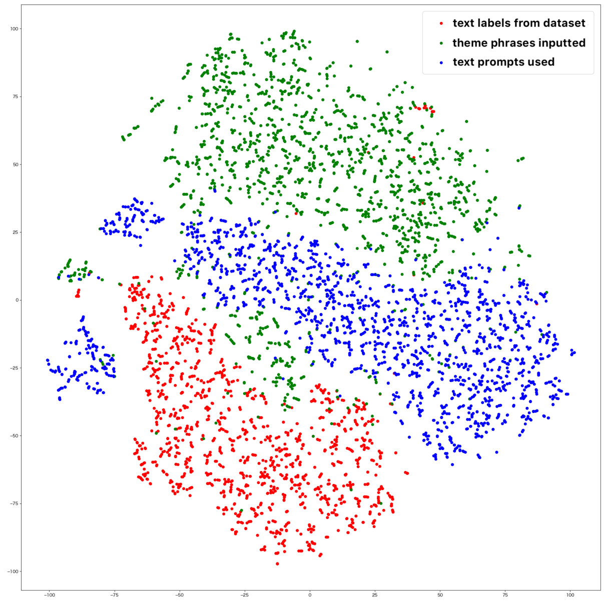

Figure 4. Visualization of the representation vectors of the theme phrases inputted by the users, the text prompts computationally derived from them, and the text labels in the training dataset.

5.2 Journey of iterative exploration

We also investigated how the users interacted with generated results produced by IteraTTA. We analyzed the in-

[3] Its English version is currently available at http://iteratta.duckdns.org/, and readers can try it on their Web browsers (Google Chrome is recommended).

[4] For the text labels, we extracted labels containing "music" from AudioCaps [13].

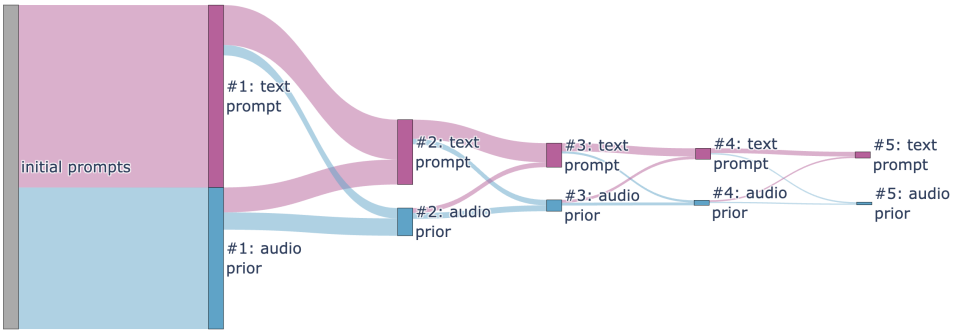

Figure 5. Visualization of how the users utilized the dual-sided exploration of IteraTTA.

teraction log of the service and obtained Figure 5. While some users just tried the exploration feature once, we found that others made iterative use of the feature, alternating between providing text prompts and audio priors. Interestingly, one user repeated this refinement process 32 times, specifying text prompts 14 times and audio priors 18 times before sharing their final result on Twitter. These points imply that our design, which enables dual-sided iterative exploration, helped the users effectively utilize the text-to-audio model.

5.3 Unleashing the creativity of novice users

We lastly analyzed the users' responses to the feedback form, which received 33 responses in total. Overall, most of them expressed their affirmative experiences with the text-to-audio music creation processes, like:

> *"It was a very interesting trial. I can interact with it throughout the day."*
>
> *"In my personal opinion, it can be used as a source of sampling materials and an idea generator. As a person who usually composes music, I never had any negative feelings about composing from text using this. It is wonderful."*

The latter comment suggests that the features of IteraTTA prepared for novice users can also benefit experienced users in different ways.

It is also notable that the users left comments implying the importance of the design requirements discussed in Section 3, such as how they enjoyed the open-ended exploration starting from loosely-specified theme phrases.

> *"It was fun to encounter songs that fit the theme I provided but I had never heard before."*
>
> *"I really enjoyed the points that I could take advantage of ChatGPT's ability to associate and verbalize even seemingly unconnected ideas, which allowed me to provide crazy theme phrases that would not be understood by a human. I also learned a lot about how to describe songs by looking at the derived text prompts."*

Interestingly, in the form, some users left a successful prompt that they reached after exploration:

> *"I would like to report that including a phrase of 'simple progression' or limiting the number of tracks yielded stabilized music audios, like: 'Ideal harmonious song: balanced instrumentation, band sound, simple chord progressions, rhythmic drum patterns, catchy pop melody, up to 12 tracks.' "*
>
> *"Adding 'clear sound quality' produces less noisy audios."*

It is surprising that, even though we provided no explicit description of the behavior of text-to-audio models, the users were able to gain such knowledge by themselves through the iterative exploration with IteraTTA. While such *prompt modifiers* (also known as *quality boosters*) [64] that influence results in a specific way have been discovered for text-to-image models in a community-driven manner [17, 64], the above comments would be the first examples for text-to-audio models, to the best of our knowledge. We assume that this is a manifestation of users' creativity in text-to-audio music generation processes and would be hard to derive without IteraTTA.

6. CONCLUSION

This paper introduces IteraTTA, an interface specifically designed for supporting novice users in their text-to-audio music generation processes. Its design is guided by two main principles, providing a) computational guidance for constructing initial prompts and b) dual-sided iterative exploration of text prompts and audio priors. The former can help novice users translate their loosely-specified goals into text prompts, which serve as starting points for exploration, even if they do not have rich artistic vocabularies. The latter is important for enabling them to comprehend the space of possible results and gradually refine their goals. To examine how diverse users utilize IteraTTA in their creative processes, we deployed it as a publicly-available Web service and analyzed users' behaviors, which highlight the importance of these design considerations in supporting the users' creativity. Importantly, these principles are applicable not only to the specific text-to-audio model but to other models, including those to be proposed in the near future. We believe that this paper can serve as a foundation for enabling novice users to benefit from state-of-the-art models in the MIR community.

7. ACKNOWLEDGEMENT

This work was supported in part by JSPS KAKENHI Grant Number JP21J20353, JST ACT-X Grant Number JPMJAX200R, and JST CREST Grant Number JPMJCR20D4, Japan.

8. REFERENCES

[1] G. Franceschelli and M. Musolesi, "Creativity and machine learning: A survey," *arXiv*, vol. abs/2104.02726, 2021.

[2] M. J. Muller, L. B. Chilton, A. Kantosalo, C. P. Martin, and G. Walsh, "Proceedings of the GenAICHI workshop: Generative AI and HCI," in *Extended Abstracts of the 2022 ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2022, pp. 110:1–110:7.

[3] F. Carnovalini and A. Rodà, "Computational creativity and music generation systems: An introduction to the state of the art," *Frontiers in Artificial Intelligence*, vol. 3, p. 14, 2020.

[4] M. Rohrmeier, "On creativity, music's AI completeness, and four challenges for artificial musical creativity," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, pp. 50–66, 2022.

[5] H. H. Tan and D. Herremans, "Music FaderNets: Controllable music generation based on high-level features via low-level feature modelling," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 109–116.

[6] K. Chen, C. Wang, T. Berg-Kirkpatrick, and S. Dubnov, "Music SketchNet: Controllable music generation via factorized representations of pitch and rhythm," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 77–84.

[7] Z. Wang, D. Wang, Y. Zhang, and G. Xia, "Learning interpretable representation for controllable polyphonic music generation," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 662–669.

[8] S. Dai, Z. Jin, C. Gomes, and R. B. Dannenberg, "Controllable deep melody generation via hierarchical music structure representation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021.

[9] Z. Wang and G. Xia, "MuseBERT: Pre-training music representation for music understanding and controllable generation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 722–729.

[10] A. Agostinelli, T. I. Denk, Z. Borsos, J. H. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. H. Frank, "MusicLM: Generating music from text," *arXiv*, vol. abs/2301.11325, 2023.

[11] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. P. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," *arXiv*, vol. abs/2301.12503, 2023.

[12] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 776–780.

[13] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 2019, pp. 119–132.

[14] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 736–740.

[15] K. Swanwick, *Musical Knowledge: Intuition, analysis and music education*. London, UK: Routledge, 2002.

[16] J. E. Gromko, "Perceptual differences between expert and novice music listeners: A multidimensional scaling analysis," *Psychology of Music*, vol. 21, no. 1, pp. 34–47, 1993.

[17] J. Oppenlaender, R. Linder, and J. Silvennoinen, "Prompting AI art: An investigation into the creative skill of prompt engineering," *arXiv*, vol. abs/2303.13534, 2023.

[18] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[19] K. Andersen and P. Knees, "Conversations with expert users in music retrieval and research challenges for creative MIR," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*. ISMIR, 2016, pp. 122–128.

[20] C. Roads, "Research in music and artificial intelligence," *ACM Computing Surveys*, vol. 17, no. 2, pp. 163–190, 1985.

[21] J. D. Fernández and F. J. Vico, "AI methods in algorithmic composition: A comprehensive survey," *Journal of Artificial Intelligence Research*, vol. 48, pp. 513–582, 2013.

[22] C. Liu and C. Ting, "Computational intelligence in music composition: A survey," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 1, no. 1, pp. 2–15, 2017.

[23] J. Briot and F. Pachet, "Deep learning for music generation: challenges and directions," *Neural Computing and Applications*, vol. 32, no. 4, pp. 981–993, 2020.

[24] E. Deruty, M. Grachten, S. Lattner, J. Nistal, and C. Aouameur, "On the development and practice of AI technology for contemporary popular music production," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, p. 35, 2022.

[25] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," *arXiv*, vol. abs/2011.06801, 2020.

[26] L. Yang, S. Chou, and Y. Yang, "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation," in *Proceedings of the 18th International Society for Music Information Retrieval Conference*. ISMIR, 2017, pp. 324–331.

[27] H. Dong and Y. Yang, "Convolutional generative adversarial networks with binary neurons for polyphonic music generation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. ISMIR, 2018, pp. 190–196.

[28] C. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music Transformer: Generating music with long-term structure," in *Proceedings of the 7th International Conference on Learning Representations*. OpenReview.net, 2019.

[29] Y. Huang and Y. Yang, "Pop Music Transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 1180–1188.

[30] W. Hsiao, J. Liu, Y. Yeh, and Y. Yang, "Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI Press, 2021, pp. 178–186.

[31] G. Mittal, J. H. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 468–475.

[32] B. Yu, P. Lu, R. Wang, W. Hu, X. Tan, W. Ye, S. Zhang, T. Qin, and T. Liu, "Museformer: Transformer with fine- and coarse-grained attention for music generation," in *Proceedings of the 36th Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2022, pp. 1376–1388.

[33] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv*, vol. abs/2005.00341, 2020.

[34] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv*, vol. abs/2111.05011, 2021.

[35] T. Hung, B. Chen, Y. Yeh, and Y. Yang, "A benchmarking initiative for audio-domain music generation using the FreeSound loop dataset," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 310–317.

[36] M. Pasini and J. Schlüter, "Musika! Fast infinite waveform music generation," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, 2022, pp. 543–550.

[37] T. Akama, "Connective fusion: Learning transformational joining of sequences with application to melody creation," in *Proceedings of the 21th International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 46–53.

[38] H. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y. Yang, "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, 2021, pp. 318–325.

[39] P. Neves, J. Fornari, and J. B. Florindo, "Generating music with sentiment using Transformer-GANs," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, 2022, pp. 717–725.

[40] K. Wang and J. V. Nickerson, "A literature review on individual creativity support systems," *Computers in Human Behavior*, vol. 74, pp. 139–151, 2017.

[41] C. A. Huang, H. V. Koops, E. Newton-Rex, M. Dinculescu, and C. Cai, "Human-AI co-creation in songwriting," in *Proceedings of the 21th International Society for Music Information Retrieval Conference*. ISMIR, 2020, pp. 708–716.

[42] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[43] J. Heer, "Agency plus automation: Designing artificial intelligence into interactive systems," *Proceedings of the National Academy of Sciences*, vol. 116, no. 6, pp. 1844–1850, 2019.

[44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv*, vol. abs/1907.11692, 2019.

[45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 10 674–10 685.

[46] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in *Proceedings of the 17th European Conference on Computer Vision*. Springer, 2022, pp. 89–106.

[47] I. Simon, D. Morris, and S. Basu, "MySong: automatic accompaniment generation for vocal melodies," in *Proceedings of the 2008 ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 725–734.

[48] C. A. Huang, C. Hawthorne, A. Roberts, M. Dinculescu, J. Wexler, L. Hong, and J. Howcroft, "The Bach doodle: Approachable music composition with machine learning at scale," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. ISMIR, 2019, pp. 793–800.

[49] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, "Novice-AI music co-creation via AI-steering tools for deep generative models," in *Proceedings of the 2020 ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 610:1–610:13.

[50] S. Rau, F. Heyen, S. Wagner, and M. Sedlmair, "Visualization for AI-assisted composing," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, 2022, pp. 151–159.

[51] Y. Zhang, G. Xia, M. Levy, and S. Dixon, "COSMIC: A conversational interface for human-AI music co-creation," in *Proceedings of the 21th International Conference on New Interfaces for Musical Expression*. nime.org, 2021.

[52] Y. Zhou, Y. Koyama, M. Goto, and T. Igarashi, "Generative melody composition with human-in-the-loop bayesian optimization," in *Proceedings of the 2020 Joint Conference on AI Music Creativity*. DiVA.org, 2020.

[53] ——, "Interactive exploration-exploitation balancing for generative melody composition," in *Proceedings of the 26th International Conference on Intelligent User Interfaces*. ACM, 2021, pp. 43–47.

[54] P. C. Wright and A. F. Monk, "The use of think-aloud evaluation methods in design," *ACM SIGCHI Bulletin*, vol. 23, no. 1, pp. 55–57, 1991.

[55] O. Alhadreti and P. J. Mayhew, "Rethinking thinking aloud: A comparison of three think-aloud protocols," in *Proceedings of the 2018 ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 44.

[56] A. Strauss and J. Corbin, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Newbury Park, CA: Sage Publications, 1990.

[57] J. O. Talton, D. Gibson, L. Yang, P. Hanrahan, and V. Koltun, "Exploratory modeling with collaborative design spaces," *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 1–10, 2009.

[58] C. Lynch, K. D. Ashley, N. Pinkwart, and V. Aleven, "Concepts, structures, and goals: Redefining ill-definedness," *International Journal of Artificial Intelligence in Education*, vol. 19, no. 3, pp. 253—266, 2009.

[59] H. Yakura, Y. Koyama, and M. Goto, "Tool- and domain-agnostic parameterization of style transfer effects leveraging pretrained perceptual metrics," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. IJCAI, 2021, pp. 1208–1216.

[60] L. Tweedie, "Interactive visualisation artifacts: How can abstractions inform design?" in *Proceedings of the 10th BCS Conference on Human-Computer Interaction*. Cambridge University Press, 1995, pp. 247–265.

[61] M. A. Terry and E. D. Mynatt, "Side views: Persistent, on-demand previews for open-ended tasks," in *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2002, pp. 71–80.

[62] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, vol. *in press*, 2022.

[63] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[64] J. Oppenlaender, "A taxonomy of prompt modifiers for text-to-image generation," *arXiv*, vol. abs/2204.13988, 2022.