Sound Designer-Generative AI Interactions: Towards Designing Creative Support Tools for Professional Sound Designers

Purnima Kamath
purnima.kamath@u.nus.edu
Augmented Human Lab
National University of Singapore
Singapore

Fabio Morreale
f.morreale@auckland.ac.nz
School of Music
University of Auckland
New Zealand

Priambudi Lintang Bagaskara
e1101547@u.nus.edu
Augmented Human Lab
National University of Singapore
Singapore

Yize Wei
yize.wei@u.nus.edu
Department of Computer Science
National University of Singapore
Singapore

Suranga Nanayakkara
suranga@ahlab.org
Augmented Human Lab
National University of Singapore
Singapore

ABSTRACT

The practice of sound design involves creating and manipulating environmental sounds for music, films, or games. Recently, an increasing number of studies have adopted generative AI to assist in sound design co-creation. Most of these studies focus on the needs of novices, and less on the pragmatic needs of sound design practitioners. In this paper, we aim to understand how generative AI models might support sound designers in their practice. We designed two interactive generative AI models as Creative Support Tools (CSTs) and invited nine professional sound design practitioners to apply the CSTs in their practice. We conducted semi-structured interviews and reflected on the challenges and opportunities of using generative AI in mixed-initiative interfaces for sound design. We provide insights into sound designers' expectations of generative AI and highlight opportunities to situate generative AI-based tools within the design process. Finally, we discuss design considerations for human-AI interaction researchers working with audio.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Artificial intelligence**; • **Applied computing → Sound and music computing**.

KEYWORDS

Audio, Generative AI, Sound design, Creative Support Tools, Mixed-Initiative Creative Interfaces

ACM Reference Format:
Purnima Kamath, Fabio Morreale, Priambudi Lintang Bagaskara, Yize Wei, and Suranga Nanayakkara. 2024. Sound Designer-Generative AI Interactions: Towards Designing Creative Support Tools for Professional Sound Designers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3642040

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642040

1 INTRODUCTION

The practice of sound design is the creative use of sounds [89] to produce innovative music compositions and build convincing cinematic experiences for films and games [70]. This practice involves the creation of new sounds and manipulation of existing ambient sound recordings, such as a dog barking or footsteps. In the last few years, generative models for audio have been applied to creating music [1, 24, 60] and have moved from being exclusively a research endeavor to finding practical applications [4, 8, 30]. Such models are well studied for their potential to support co-creation in the human-AI interaction literature [61, 62]. And yet, despite the growing adoption of AI models as co-creation tools for music production [28], very few empirical studies exist to assess their potential to offer new possibilities to the practice of sound design and foley sound synthesis.

Our everyday sonic environment is usually composed of not just music or speech, but also a myriad of environmental sounds [89]. Often, sound designers work with environmental sounds that lack the rhythmic and harmonic structures normally found in musical compositions. Further, the parameters used for synthesizing such sounds are different from musical sounds. For instance, when synthesizing footstep sounds, sound designers are more likely to be interested in manipulating object and material properties (such as type of floor, shoes, etc.) than acoustic features such as pitch or loudness, which are typically associated with music. Similarly, AI models trained on environmental sounds differ from AI models for music in the way they are controlled or steered [72] using either musical or material attributes during generation. Thus, AI-based tools used to assist sound designers need to be studied through specifically designed studies.

Most human-AI interaction studies for audio focus on the applicability of such steerable interfaces to empower novice users in their creative goals [7, 36, 61, 62, 102]. Expert sound design practitioners spend years developing their creative design process and building inventories of sounds to apply in their next design project [89]. As such, their needs, expectations, and ways of working with AI-based

tools necessarily differ from those of novices'. Thus, in this paper, we aim to explore: *How can generative AI-based co-creation tools assist expert sound designers in their creative practice?*

We developed two interactive generative AI models as Creative Support Tools (CSTs) [35, 80] to explore the potential of this technology in assisting sound designers. As in [49], we use an experimental design strategy of deploying interfaces in real-world contexts to provoke discussions and answer research questions. We deployed our CSTs with sound designers to gather information about their expectations of AI, and the current challenges and opportunities for generative AI in their practice. Further, we captured the designer's interpretations of AI by designing the interactivity with our CSTs using an element of ambiguity [37]. While we developed two CSTs in this study, we did not aim to compare them with each other. Instead, we aimed to provide designers with two unique ways of interacting with AI-based tools to gather their reflections [34, 49] on using those tools.

We introduced our CSTs to nine professional sound designers and asked them to apply them in a creative endeavor in their practice. We conducted semi-structured interviews with the participants to reflect on their creative goals and the sounds they created using the CSTs. We gained three key insights through inductive reflexive thematic analysis [9] of the interviews:

- First, we outline *an AI-assisted sound design process* where we find how sound designers situate AI-based tools within their design process in practice. While performing creative tasks, we found that sound designers used AI models for performing fast iterations to create novel sounds and as an alternative to manual field recording activities. They also used such sounds as layers to give the perception of plausibility to unreal sounds.
- Next, we found how sound designers *worked with unpredictability and ambiguity* and developed an intuition for interacting and controlling AI-generated sounds. We also found that designers often realized or understood failure modes in AI-generated output and worked towards ways of using ambiguity in their sound design.
- Lastly, we furthered our understanding of *sounds designers' expectations of generative AI* to build convincing cinematic experiences, in terms of creator agency and owning the creative process

In summary, our contributions are three-fold: (1) we developed a novel understanding of generative AI in supporting creative exploration for the practice of sound design; (2) we developed two AI-based CSTs for future studies on using audio generative AI as a tool for sound designers and (3) we offered five design recommendations for future human-AI interaction research for sound design.

2 BACKGROUND & RELATED WORK

2.1 Sound Design

Sound design is a multi-faceted practice that is both highly technical and artistic in nature and involves creating 'new' sounds. Susini et al., [89] define 'new' sounds as those that cannot be found in existing sound databases, or recorded sounds that cannot be used in a given context without being manipulated or modified. Sound design is the deliberate use of such sounds to create an 'atmosphere', mood, or feeling [70] in music composition or other media. Typically, sound designers focus on working out the sonic details or timbres (tone or color of the sound) required to enrich or complement the visual information presented in films and games [19]. They also focus on communicating additional non-verbal information through interactions in games or product design [89].

Previously, Wallas [95] proposed a generalized creativity model which consisted of 4 phases - preparation, incubation, illumination, and verification. Similarly, for the specific purpose of sound design, Susini et al. [89] proposed a model which involved three discrete successive stages: *Analysis* assisted by *Exploration*, *Creation*, and *Validation* [58, 89] with the last two being set in an iterative loop until the sound converges towards an optimized solution [47]. The *Analysis* stage is a research-focused phase, where designers are involved in understanding the perceptual requirements of the project using their own knowledge and background in psychoacoustics and sound cognition. It also involves the purposeful exploration of a large inventory of existing sounds as well as field recording (recording outside of the studio) of new sounds. In the *Creation* stage, designers manipulate the sounds or synthetically create new sounds in line with the specifications from the *Analysis* stage. This stage may also layer together various sound samples to create montages as a final artifact. The final stage consists of *Validating* the sound specimens created either informally based on the designer's intuition or more formally using listening tests (especially while designing sounds for products [58]).

Throughout the sound design process, designers need to employ different modes of working - as a researcher during *Analysis* and exploration phase, as a programmer or employing their tools-based expertise during the *Creation* phase, and as a qualitative researcher or tester during the *Validation* phase. Through these phases, they also employ different listening techniques such as *causal*, *semantic*, or *reduced* listening [19]. Such listening techniques help designers to associate sounds to sources (*causal*), associate information or meaning (*semantic*) to them, or focus on fine-grained timbre-specific details of the sound (*reduced*) when listening. These modes of working help them develop '*Sonic Vocabularies*' [47] and '*Sound Palettes*' for various current and future projects.

In [64], Lubart suggests that computers can partner with humans as an enabler, a guide, or a partner or colleague. Similarly, in recent human-AI interaction literature, Weisz et al., [96] suggest that users may view an AI agent as filling the role of an assistant, partner, or collaborator. And that establishing the role of AI in a user's workflow will help users understand how to interact with it effectively. Thus, in this paper, we aim to situate AI-based Creative Support Tools in the context of the sound design process and a sound designer's way of working.

2.2 AI-based creativity support tools

AI algorithms have enabled the building of Creative Support Tools (CSTs) [35, 80] that are either fully autonomous or support co-creation as Mixed-Initiative Creative Interfaces (MICIs) [29, 86]. Interactive CSTs have been studied for their co-creation capabilities in visual arts [17, 27, 43, 53, 71], in writing [14, 23], in fashion [50], in UX design and engineering [38, 54, 63, 67], and in new musical

interface design [91]. In the field of audio, machine learning models have long been used for creating music [32], from established tools such as Wekinator [33] to the more recent GAN-based music performance art [90]. Further co-creation tools for composing music have been studied for novice co-creation [36, 61, 62, 66, 102] as well as with expert musicians [7, 45, 88]. By co-creation, we allude to the human-centered AI's [81] ability to leverage a trained prior to empowering human creators with novel means to generate creative artifacts. In the sound design space, Scurto et al. [77] developed tools based on reinforcement learning algorithms and studied user exploration behaviors for the generated high-dimensional parameter spaces. We take inspiration and further extend their work to situate generative audio AI models in a typical sound design process and explore designers' needs and expectations of such co-creation tools.

In [43], Hertzmann argues that *"all art algorithms, including methods based on machine learning, are tools for artists; they are not themselves artists"*. In that light, we position the interfaces in our study as AI-based MICIs, where a user expresses their intent through a set of control parameters and the AI establishes its initiative and agency by generating an output based on its trained prior. Previously, researchers argued that unpredictability and non-determinism are detrimental to the user experience of an AI system [5]. Recently, Caramiaux et al. [17] showed that such emergent behavior is embraced rather than considered a limitation in the domain of AI-generated visual arts. We take inspiration from their work to explore aspects of such non-determinism in audio AI in the specific context of sound design.

2.3 Interactive Generative AI models for Audio

Currently, a multitude of AI architectures exists for generatively modeling environmental sounds. Each architecture solves a certain set of problems and has its own limitations. For instance, while autoregressive architectures, models that predict future values based on past values, such as Recurrent Neural Networks [48, 99], WaveNet [92], or Transformers [94], trained on raw audio, can generate sounds of indefinite duration, the time taken for their responses are usually large which makes their adoption in practice difficult. On the other hand, models based on Generative Adversarial Network (GAN) [39] are responsive but generate samples of a pre-defined duration (usually a few seconds in length). Such models are expressive in their ability to allow the generation of novel sounds or morphs [41] as compared to autoregressive architectures [100]. Diffusion-based models [44, 55, 84, 85, 101] have recently become popular as an alternative to GANs. Although such architectures can generate better-quality sounds than GANs, GAN-based architectures are currently able to generate sounds faster and in real-time than diffusion models [97]. Further, a type of GAN architecture called StyleGAN [52] provides significant improvements over other GAN architectures. Thus, in this paper, we focus on using StyleGANs to develop our CSTs for sound design.

Broadly, two approaches exist for controlling generation from AI models. One where the AI models are trained on labeled datasets in a supervised way [40, pg. 137]. For such models, generation is controlled using pre-defined labels. Another approach is where AI models are trained on unlabelled datasets, and controllability is inferred using unsupervised methods [40, pg. 142]. This approach is especially useful for environmental sounds as they can be recorded easily 'in-the-wild', but it is difficult to reliably annotate them with labels. StyleGANs [52] can be trained on such large unlabeled environmental sound datasets to generate an expressive high-dimensional learned representation called a "latent space". This latent space can be used to search, generate, and manipulate new sounds. Recently, researchers have been working at the intersection of explainable AI (XAI) and arts to explore novel ways to explore such latent spaces for creative endeavors [13, 34]. Various such algorithms exist to facilitate the exploration of this high-dimensional latent space in a human-understandable and unsupervised way. For instance, in [51], the authors developed an Example-Based Framework (EBF) to search or query the latent space of a pre-trained GAN using synthetically generated sounds. Further, in [79], the authors developed a Semantic Factorization algorithm (SeFa) to find vectors for control in the latent space of the model, which can be used to manipulate semantically meaningful attributes on the sounds. In this paper, we used EBF and SeFa to interact with our underlying StyleGAN model to control the generation of sounds in our CSTs.

3 AUDIO GENERATIVE AI CST DESIGN

We designed and implemented two generative AI audio Creative Support Tools (CSTs) for co-creation in this study. We designed the interactivity for our interfaces based on the principles for AI controllability outlined in Weisz et al. [96]: by using (1) domain-specific controls (for *interface-1*), and (2) technology-specific controls (for *interface-2*). Domain-specific controls make use of audio descriptors or acoustic parameters to control the generation from an AI model. Technology-specific controls, on the other hand, are generic controls that depend on the generative algorithm and are not necessarily related to the audio domain. Such technology-specific controls allow users to perform manipulations or edits directly in the latent space of the generative model and are typically effective in making changes to the semantic attributes of a sound. For both interfaces, we adopted an interaction pattern of turn-taking [76], where a user makes modifications to the control parameters to interact with the underlying AI model and the AI responds based on its trained prior.

Both interfaces used the same underlying StyleGAN model and differed only in how the generation was controlled. Further, both interfaces provided opportunities to interact with two StyleGANs - (1) one trained on a dataset of 'Hits & Scratches' called the *Greatest Hits Dataset* [73], and (2) another trained on a dataset of 'Environmental Sounds' from the *DCASE 2023 Foley Sound Synthesis Challenge* [21]. Using the 'Hits & Scratches' model, the sound designers could generate and explore a small set of timbres related to the impact sounds made by a drumstick hitting various hard and soft surfaces. Using the 'Environmental Sounds' model, the sound designers could generate and explore more complex timbres and sounds such as dog barks, footsteps, gunshots, motor vehicles, rain, and keyboard clicks. Further, on both interfaces, we added some preset sound configurations, which participants could test during the study. These presets included parameter settings for timbres such as impact sounds on hard and soft surfaces or environmental sounds such as a medium-sized dog barking.

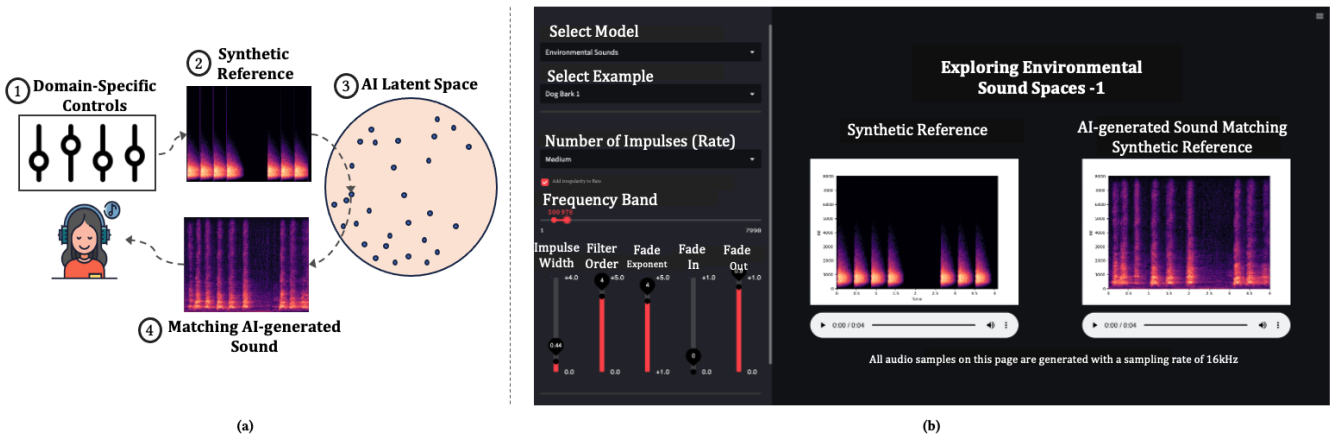

Figure 1: A conceptual diagram (a), and screenshot (b) of interface-1. (a) A sound designer can use the domain-specific controls from ① to generate a synthetic reference sound seen in ②. This synthetic reference sound is used to "query" or "search" the latent space of an AI model shown in ③ to generate a matching AI-generated sound in ④. (b) The screenshot shows the placement of the controls and the synthetic and generated sounds as viewed by the designer on the web interface. Please see Appendix A.2 for a link to a Google Colaboratory version of this interface, and A.4 for image attributions.

All underlying AI models were built and trained using Pytorch [74] and were running on a single RTX 3090 GPU. The interfaces were built as web-based technologies such as Streamlit[1] and ReactJS[2] to run on web browsers for ease of access. Please see appendix A.2 for architecture and implementation details for both interfaces.

3.1 Interface-1 - Using domain-specific controls

For *interface-1*, we employed the use of domain-specific controls [96] based on acoustic parameters such as frequency band, impulse width, fade-in, fade-out, etc. to guide the generation of the sounds. For this interface we use the EBF framework outlined in [51]. In EBF, a set of domain-specific controls is used to create a synthetic sound using signal processing techniques. This sound is then used to "query" or "search" the latent space of the StyleGAN for a matching, AI-generated sound. A conceptual diagram and screenshot of this interface are shown in Figure 1. A user of this interface conveys their ideas to the AI model by designing a synthetic sound. The AI model in turn uses the synthetic sound to search and generate a matching, more realistic sound. The resulting audio for both the synthetic reference as well as the AI-generated sounds is displayed on the webpage. Additionally, we provided visual feedback to the users by displaying the spectrogram for each sound along with the audio on the webpage. We include this spectrogram visualization to allow the participants to focus on the spectromorphology of the sounds [82], or how the frequencies in the sound change or morph over time.

While we designed this interface to provide opportunities for reflection [34, 37] by giving greater flexibility in generating multiple types of synthetic sounds, not all synthetic references resulted in meaningfully matching AI-generated sounds. This unpredictability in the AI-generated sounds is due to the limitations of the training data used to train the GAN. We allowed this unpredictability on this interface by design to gather our participants' intuition about AI limitations.

3.2 Interface-2 - Using technology-specific controls

For *interface-2*, we employed the use of technology-specific controls [96] based on the SeFa algorithm outlined in [79]. In SeFa, dimensions for controlling generation are extracted by performing an eigendecomposition of the learned weights of the StyleGAN. That is, using eigendecomposition, the weights matrix of the StyleGAN are factorized into basis vectors which can then be used to perform latent space manipulations to control semantic audio descriptors on a sound. Such semantic dimensions are usually unlabeled and are typically open to user interpretation of them. Users usually interpret each semantic dimension by performing and observing a few edits made by changing a dimension on the sound. We chose the top 10 dimensions (top 10 eigenvalues after eigendecomposition, see appendix A.2.2) found by the algorithm to perform sound edits on this interface. A conceptual diagram and screenshot of this interface are shown in Figure 2. As for interface-1, we displayed the spectrogram along with the resulting audio on this interface.

We designed this interface to provide opportunities for reflection [34] by leaving the dimensions unlabeled. We allowed the designers to interpret this ambiguity in the dimensions based on their intuition.

4 USER STUDY

4.1 Participants

With ethics approval obtained from the University, we recruited nine professional sound design practitioners (six male, two female, one preferred not to say) for this study through snowball sampling. We used this sampling strategy to reach not just academic, but also professional sound designers working in the industry. Starting with

[1]https://streamlit.io/
[2]https://react.dev/

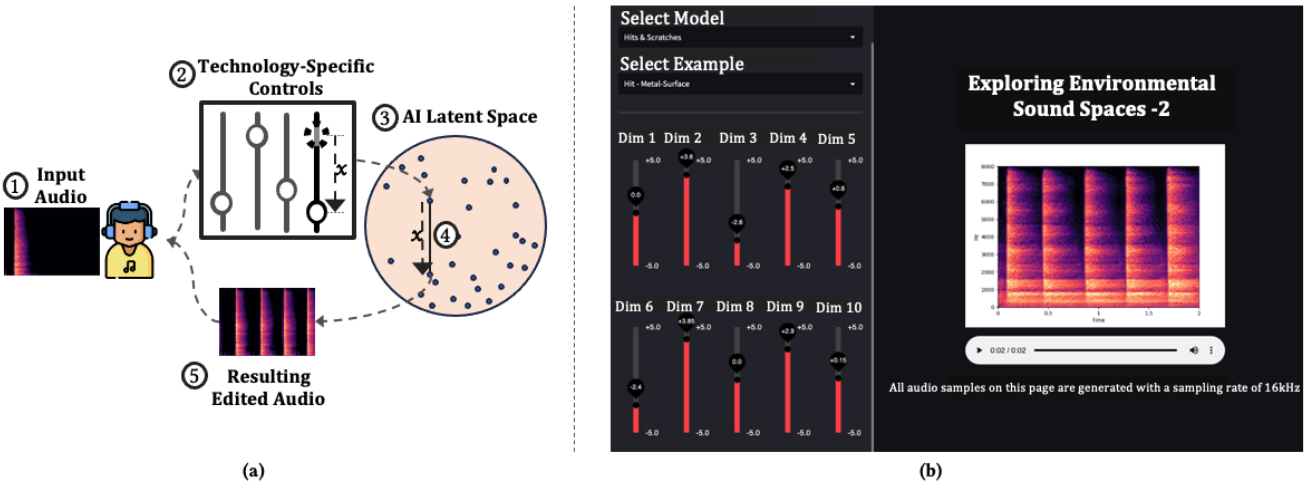

Figure 2: A conceptual diagram (a), and screenshot (b) of interface-2. (a) To edit the audio in ① such that the number of impacts in the sound increases, a sound designer can use the technology-specific controls extracted from the latent space of a StyleGAN shown in ② to perform direct latent space manipulation shown in ③ and ④, resulting in the edited audio sample in ⑤. (b) The screenshot shows the placement of the controls and the generated sounds as viewed by the designer on the web interface. Please see Appendix A.2 for a link to a Google Colaboratory version of this interface, and A.4 for image attributions.

the authors' existing network, we then asked individual participants whether they knew of other practitioners interested in participating in our study. In our email, we indicated the study would take at least 1.5 hours to complete. Our sample size was thus pragmatic based on the number of sound designers who were willing to invest the time in this study. Participants had diverse backgrounds in sound design, from designing sounds for products, movies, music, and games to creating sound for data sonification projects (Table 1). The median self-reported years of experience in sound design was 10 years (Min = 3 years, Max = 48 years). They were offered USD45 gift cards as a token of appreciation for their time in the study.

4.2 Procedure

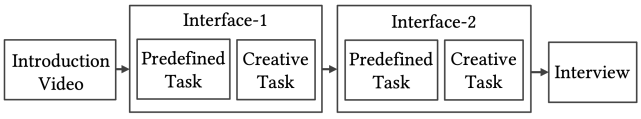

Figure 3: Overview of the study procedure

An overview of the procedure is shown in Figure 3. The participants were sent a link to a web page outlining the task instructions[3]. This web page included a 3-minute introductory video explaining the tasks and providing a brief overview of the interfaces. To minimize any order effects, participants were randomly assigned into two groups. The first group attempted the tasks with interface-1 first before interface-2. The second group performed the tasks in the reverse order.

To familiarize participants with our interfaces, we first asked them to complete a short close-ended predefined task. Subsequently, we asked them to complete an open-ended creative task to generate sounds they might use in their own practice or performance. As our participants were located in different parts of the world, they were asked to perform these tasks at their own pace and time and record their screen activity when performing the open-ended creative task. This approach was adapted from the video-cue recall method [15, 16] from the interactive arts literature for our purpose.

We subsequently conducted a semi-structured interview to gauge the participants' experience and feedback on the generative AI interfaces. Our server logs indicated that overall, the participants spent a median of 46.28 minutes (Min=24.11 minutes, Max=2 hours, 31.6 minutes, SD=44.46 minutes) exploring and familiarizing themselves with the interfaces. As instructed, participants recorded their screen activity when performing their open-ended creative tasks. For these creative tasks, the median screen recordings for each interface were 2.44 minutes long (Min=1 minute, Max=20.25 minutes, SD=5.36 minutes). We asked the participants to send us their screen recordings in advance of the interview. The interviewer watched the screen recordings before conducting the interview and highlighted parts of the recording where participants employed different exploration strategies when using the interfaces. During the interview, we discussed the participants' creative goals using the highlighted parts of the recordings as discussion prompts. The recordings were used as discussion prompts only and not as data for analysis in themselves. All interviews were conducted remotely and lasted for a median of 40 minutes (Min=32 minutes, Max=60 minutes, SD=10.19 minutes). Please see Appendix A.1 for the interview questions.

[3]See link on our webpage - https://pkamath2.github.io/chi2024-resources/#task-instructions

Table 1: Participant Details

ID	Country	Experience	Description of Sound Design Experience
P1	New Zealand	8 years	Sound design for visual media such as short films, documentaries, and games. Using sound as an aid and embellishment to story-telling. Experience in recording and mixing music. Undertaken audio post-production work.
P2	Spain	48 years	Sound designer and electronic music composer. Focused on audio perception and programmatic ways of creating sound. Worked on programmable synthesizers and libraries for various platforms. Educator for sound art and design. Currently focused on audio AI research.
P3	Hong Kong SAR (China)	9 years	Sound design for movies and animated films. Audio post-production for TV programs. Experience recording and mixing music, and foley sound effects.
P4	New Zealand	7 years	Original sound creation and implementation for online and theatrical films, games, music production, and live performances. Field recording. Post-production work includes dialogue editing, sound mixing, audio restoration, and foley mixer. Educator for sound design.
P5	Netherlands	20 years	Designed sound to build brand experiences for various international brands and airport authorities. Designed "sonic identities" for brands ranging from sound installations for their public spaces as well as designing product sounds. E.g., the sound of a car's engine, doors opening or closing, etc. Focussing also on data sonification projects.
P6	Italy	3 years	Sound design for vehicle or gardening simulation video games working directly with environmental soundscapes. Designing quad ambiance and sound effects, and implementing them in the game engine.
P7	Germany	10 years	Sound designer and composer. Designed sounds for over 40 games. Also worked on sound design for films as well as some movie trailers.
P8	Singapore	10 years	Electroacoustic music composer using Ableton Live and FL Studio.
P9	Singapore	47 years	Music composition for ambient/rock and experimental/avant-garde genres. Worked for theatre and other projects that are in between sound design and music. Co-leader for a desktop Foley system. Also writing music software.

4.3 Data Analysis

Due to the exploratory nature of this work, we chose an inductive, reflexive thematic analysis (TA) approach [9] for analyzing the interview transcripts. One author conducted all interviews. Two authors (including the interviewer) collaboratively analyzed the data using a bottom-up approach. We first familiarized ourselves with the transcripts by individually and independently reading them at least twice. We then coded the transcripts with quotes relevant to our research objectives. Next, we collaboratively combined and refined our codes using Atlas.TI[4]. As recommended in [10], we use a semantic coding strategy during our coding process where each code captures a semantic observation. For instance, a quote from a participant such as "some randomness (in the AI-generated output) is always refreshing" is considered as one code. Quotes from other participants making similar observations may also be tagged to the same code. This code, amongst similar other codes, is then organized under a theme such as "Non-determinism assists creativity". Through this process, we iteratively refined and identified 76 codes. We use affinity diagramming to assist us in collaboratively organizing the codes into 12 themes. These themes are organized under the 3 sections in the results section below.

In reflexive TA, meaning is not *"excavated"* [11] from the data, but is subjectively generated through a researcher's interpretation of the data [9]. This nature of the analysis makes it difficult to formalize a sample size or define data saturation (or the minimum number of participants needed before stopping data collection) [11]. Thus, instead of defining data saturation for this study, we resorted to deliberately seeking a varied group of participants based on their geographic location, background in sound design, and number of years of experience in sound design. With this, we tried to gather diverse views and opinions of AI during our study.

5 THEMATIC ANALYSIS FINDINGS

In the following three subsections, we organized the themes from our inductive, reflexive thematic analysis into three meta-themes: (1) An AI-assisted sound design process; (2) Working with unpredictability and ambiguity; and (3) Sound designers' expectations of AI for sound design

5.1 An AI-assisted sound design process

5.1.1 ***Fast iterative exploration.*** Sound designers are always looking for new sounds to use in their work. *"Like if you're working on a sci-fi game, then you can't just use run-of-the-mill sounds. And just so people are always looking for new sounds, like a new palette so to speak"(P1).* Some commonly shared frustrations our participants observed in their current design process were around the manual processes of creating new sounds on tight deadlines or low budgets. Creating and manipulating new sounds takes time and it can be frustrating as *"a lot of back and forth happens when someone (a client) has something on their mind that they can't verbalize and then you're trying to figure out what they want"(P1).* In such cases, being able to quickly and iteratively create novel sound samples using AI is beneficial.

> ***P1****: "It's really useful to be able to go through 20 iterations in less than half the time that it would take me to do it in the traditional way. And then because you can adjust so many parameters so quickly, then you're not stopping and changing things. You're not editing waveforms, you're not changing plugins. So I think it is really useful [...] I think people tend to overstate what creativity is. But to me personally, it is to be able to go through a lot of things quickly and to select the right bit of sound for that purpose."*

5.1.2 ***An alternative source to field recording.*** Often, sound designers sourced new sounds by field recording them and further processing them to develop new sound palettes. Such manual recording activities can be frustrating as they cannot always control a recording situation. *"You can't tell everyone in a city 'Be quiet for a second. I need to record this thing"'(P4).* Typically, a 5-second recorded audio takes a couple of hours to clean, denoise, and process before use. In such cases, AI-generated sounds can be considered as a suitable and convenient alternative for *"finding interesting source material" (P7).*

> ***P4****: "Most likely it would be I spend a day with the interface making a bunch of sounds and I just record all of them. I'd delete the ones that I don't think will be useful and I'd keep all the rest [. . .] I'd almost treat this like field recording in a sense, but instead of me actually going outside to record it, I am going into this interface to capture it."*

5.1.3 ***Creating unreal but tangible sound palettes.*** The bulk of sound in a film is usually added in post-production [70]. Sound designers typically develop and use a custom palette of sound effects for each film [98]. *"In sci-fi movies [. . .] we want to give people the kind of 'metal' feelings. That this world is made from science, and not really an actual world. To feel that it's a different world compared to my living world" (P3).* Thus, designers are often on the lookout for unreal, but plausible-sounding sound elements that assist in building immersive experiences for the consumers of such media. Using AI-assisted sound design tools in this study, designers were able to create such fantastical or alien, but tangible sounding palettes.

> ***P4****: "Obviously you can make stuff like this in a synthesizer, but the problem is it sounds like a synthesizer, it doesn't sound real.[. . .] And while this (AI-generated sound) doesn't sound like something that's real, because it's in some way based on something that is a real recording, it still has a kind of tangible quality to it. And that's kind of what the value is. You can make synthetic sounds that still sound somewhat like there's a real object doing it."*

Although the models used in this study were not trained with the goal of generating unreal sounds, the interactivity encoded in them enabled the designers to generate such sound palettes. Five designers*(P1, P3, P4, P5, P7)* noted that the generative AI tools were better used for generating such sounds rather than replicating real-world recordings.

[4]https://atlasti.com/

> ***P7**: "We are always on the hunt for those kinds of elements where we can layer something that actually exists with something that does not exist to enhance immersion for the consumer. Those are the elements that are more interesting for me personally. If I want to have the recording of a falling tree, I can just go out and record it. I don't need a tool for that."*

Further, six sound designers *(P1, P3, P4, P5, P7, P8)* we interviewed said that they rarely used sounds from their own libraries or external databases as-is in their projects. They usually processed the recordings through 'effect chains' (i.e., using a Digital Audio Workstation (DAW) to process sounds through a chain of effects such as adding/removing distortion, reverb, etc.) to fit the requirements of different projects. They found using the generative AI tools in the study useful as part of such effect chains. The interactivity in the tools could be used to extract textural components from various sounds, which can be used as layers to enrich other recorded or synthesized sounds.

> ***P7**: "(Describing their creative task result) For me, that would be like a sci-fi layer or that could be used in some trailers when there is something popping up. Or when a spaceship flies by. You can use that as a sweetener.*
> ***Interviewer**: What is a sweetener?*
> ***P7**: Yes, say you have a sound, but then you put something (.) on top of it like spices. And then it's like, wow! That's new!"*

5.1.4 ***Annoying, but Fun!*** Both AI-based tools in this study embodied non-determinism in ways of controlling the generated sounds by using either synthetic sound queries (interface-1) or unlabeled dimensions (interface-2). This nature of the AI-based tools was appreciated by our designers for their ability to allow exploration and serendipitous discovery of novel sounds, even when the sounds were not in line with the participant's original task goal. For instance, when performing his open-ended task with interface-2 *P2* said:

> ***P2**: "I understood that I was exploring and there was some discovery. So every once in a while you'll hear me say, 'Oh, I like that!'. Even though it wasn't necessarily exactly what I was looking for, it had something that I liked"*

Further, most designers noted that while this exploratory nature of the AI-based tool was fun, it would be annoying or frustrating to work with it on task-oriented work on a regular basis, especially on a deadline.

> ***P4**: "Well, one thing I found fun was seeing how the AI responded to the synthetic reference and how it didn't listen to me, right? So sometimes I made a change and it didn't quite reflect that and I found that interesting. But if I worked with this every day and I was on a project with a deadline and I really wanted it to listen to me, then I'd imagine it would stop being fun and it would start becoming frustrating trying to get it to do those specific things."*

5.2 Working with unpredictability and ambiguity

5.2.1 ***Exploration strategies.*** For interface-1 (domain-specific controls), the general exploration strategy we found amongst designers was following a 'broader first, then narrower' strategy. For instance, participant P5 said she would experiment broadly first, say using a wider range of frequencies, and then narrow down to the specific perceptual outcomes she had in mind by employing *reduced listening* (reduced listening, as explicated in section 2.1, is when designers concentrate on the sound for its own sake, as a sound object, independently of its causes or meaning [56]).

> ***P5**: "And as I say in the synthetic reference, this worked quite well because the sounds were like (MIMICKING THE SOUND OF A CICADA TRILLING), so I selected frequencies that are typical without too much thinking. Let's say higher frequencies. I did everything quite rough, not knowing the system and then trying to achieve this to get as closer as I could (to the goal)."*

For interface-2 (technology-specific controls), to understand the parameter space they were exploring, participants employed multiple strategies such as - (1) simply playing around with each control and observing its effect on the generated sound (*P1*, *P2*, *P5*, *P8*, and *P9*), or (2) by using a 'Systematic Change without Compounding' where the parameters are reset to the original positions first and only one parameter is changed at a time to observe or isolate its change *(P4)*, or (3) by using a 'Min-Max' strategy by observing the generated output at the minimum and maximum limits of a parameter's range *(P3, P7)*. While P3 used the 'Min-Max' strategy to clearly isolate the change made by a parameter, P7 used that strategy to see how far he could push a control to get something *"new or weird"(P7)* out of it.

Overall, we observed that the participants who approached the exploration with both interfaces systematically discovered new sounds and were generally satisfied with the exploration, even when the outcomes did not match their original goals. One participant (P6), who reportedly approached the exploration randomly and without a goal, found it difficult to get any satisfactory output and gave up performing the task. Although other participants discovered interesting new sounds from their explorations, they expressed their desire for more predictability in the controls so as to be able to use the tools regularly.

5.2.2 ***Opportunities from ambiguity.*** As outlined in Section 3, both interfaces used the same underlying AI model, but with different interactivity mechanisms governed by different levels of ambiguity, to control the generation of sounds. All designers in this study noted that while both interfaces could generate unpredictable outputs from the AI models, the controls on interface-1 (domain-specific controls) were more intuitive and comprehensible as compared to those on interface-2 (technology-specific controls). This was primarily because interface-2 had (1) unlabeled controls, and (2) a higher number of controls than those on interface-1. When using interface-2, some designers noted that the exploration seemed like a *"trial and error" (P7)*, while others *(P2, P3, P4, P5)* found that this *"lowest form of control" (P4)* gave them greater opportunities

for exploration as there were more control parameters to *"twist" (P3).*

> ***P4**: "One thing, of course, is it's less intuitive in the sense that nothing's labeled, [...] but by not giving it a name, it actually made more sense in a way, because you just see that as an abstract quality, the AI is doing something with it. So just naming them arbitrarily kind of made you pay attention to what they were actually doing more and not expecting something that it wasn't going to do. The lack of specificity makes it feel open in a different way."*

Further, when using interface-2, all designers expressed the need to be able to label the dimension based on their own preferences. Designers gravitated towards labeling the dimensions based on either semantic changes *(P1, P6, P8)* or acoustic changes *(P2, P7)* they observed in the generated output.

5.2.3 ***Modes of working with audio interfaces.*** Although designers indicated that labeling dimensions would enable them to use the interfaces better, especially when using interface-2 (technology-specific controls), two designers *(P3, P5)* reported that they relied on auditioning to understand the role of each parameter, even when using labeled controls on interface-1 (domain-specific controls). Such designers built an intuitive knowledge about the effect of each control parameter on the generated sounds and did not rely on the descriptions provided on the interface.

> ***P5**: "Usually I don't even read descriptions much. I just listen to what comes out. It's a nicer way of exploring for me. And then when I'm familiar, I can control it."*

Further, while using interface-1, we noticed five designers *(P1, P2, P3, P5, P8)* stopped listening to the synthetic sounds and focused on listening to the effect of the parameter change directly on the matching AI-generated sound itself. Reading and observing the changes on the synthetic spectrogram was sufficient for them to understand the effect of the change they made. And thus they could focus more on the effect of their changes on the AI-generated output.

> ***P2**: Since [..] the AI-generated (sound) was really what I was exploring, [..] and I can read the spectrograms well enough to know that I just didn't have to go through that intermediate step. So spectrograms were helpful in kind of building out what the goal was."*

Finally, two designers *(P3, P8)* found it easier to create atomic units of sounds, such as a single impact sound or a single dog bark. Then fixing and editing the important semantic and perceptual aspects of that single unit, and subsequently looping or repeating it in a DAW. This gave them better control of the creative process in adjusting the variability of the sounds to their liking.

> ***P3**: "I want to create the sound that is, actually can be used in my work. I think it should be one - how to say, one should sound, not (THUD THUD THUD THUD). Only one (THUD). If I need more of this, I can copy-paste (loop or repeat it in a DAW)."*

5.2.4 ***Understanding unpredictability of the response.*** As outlined in section 3.1, for interface-1 (domain-specific controls) we gave greater flexibility in generating the synthetic sounds, while not all synthetic sounds resulted in meaningfully matching AI-generated sounds. For instance, the *Greatest Hits* dataset was limited by a certain range of rate of impact (number of impact events per second). When designers tried to query the AI model for higher rates, the model generated unpredictable responses. During our interviews, we discussed the nature of the generated sounds and asked our participants if they understood the reasons behind the AI's unpredictability due to its limitations. Three participants *(P2, P4, P5)* were familiar with the idea that AI was limited by its training data. Participant *P2* had experience with building and using AI models, and participants *P4* and *P5* were familiar with popular generative models such as ChatGPT [3] or DALLE-2 [2]. Participants' prior experience with the limitations of generative AI, across different modalities, might have made it easier for them to reconcile their understanding of the failure modes in our interfaces, especially when the changes they made did not align with their expectations. For instance, while explaining the unpredictable response from interface-1, P4 said:

> *"**P4**: Often, changes in synthetic reference didn't clearly correlate to the changes in the AI-generated sound. [...] Sometimes the fade-out parameter didn't really do that much to the AI-generated sound.*
> ***Interviewer**: Can you tell me why you think that is happening?*
> ***P4**: Why? Not exactly sure why it wasn't following along exactly, but I'm guessing it's because it's trained on a certain kind of response that already has a certain type of fade-out innate in it, and so when you change the fade-out, there's only so much it can change based on what kind of input it has had."*

5.3 Sound designers' expectations of generative AI

5.3.1 ***Cinematic effect over accuracy.*** Through our interviews, we found that interviewees focussed mostly on the overall perceptual aspects of the sounds they worked with. Aspects such as where the sound originated from were not necessarily important to them. For instance, although we set up our AI CSTs to generate 'Hits & Scratches' impact sounds made by a drumstick, the sound designers used the models to create novel base sounds and *sweeteners* for footsteps *(P1)*, fantastical 'adolescent monsters' *(P3)*, trilling cicadas *(P5)*, sci-fi whooshes and flying machines *(P7)*, and as layers over percussive drum beat *(P8)*.

> ***P1**: "I think the most important thing, whether it's movies or games, is not accuracy so much, but immersion. So the footsteps that you hear in a movie, do not sound like that in the real world. Like, if you punch someone in the real world, it doesn't sound anything like what it does when Harrison Ford punches someone. The whole point (of sound design) is immersion and entertainment."*

5.3.2 ***Creative agency and ownership.*** Currently, most research in generative AI focuses on building omnipotent intelligent agents that can do it all - agents that can create art or compose music

directly instead of being an enabler for creativity. While tools with greater AI agency would work well for novice users, for sound design experts, there are more opportunities for AI as an enabler rather than being a creator in itself.

> ***P4**: "So a lot of other AI seem to be trying to replace a creator so that someone can get sounds who don't know how to make them, whereas this one seems more useful for someone who already knows how to make sounds but just wants to add to their arsenal by having another tool."*

In [89], Susini, et al. emphasized that sound design as a practice is not just concerned with generating new sounds, but is also associated with a designer-led research-oriented design process grounded in psychoacoustics and sound cognition. Although most generative AI systems focus exclusively on the generation of new sounds, they do not focus on *"what the sound should do, or what it should be" (P5)*. As such, the results from our interviews suggest that the best use of AI is as a Creative Support Tool, as a part of a larger creative process owned and controlled by the designer.

> ***P5**: "I would like to keep the ownership of the creative process. I imagine the sound as it should be because it comes from a long research [...] The creative design process is much more than making the sounds. It is more about knowing what you want and finding the right tools.[...] So if the AI is also part of the research process, it could have good ideas."*

Finally, our results indicate that AI algorithms have the technological capability to provide means for creators with novel ways of creating sound for their work which traditional signal-processing techniques cannot do. For instance, in our study, we observe two such instances where designers were able to discover novel base sounds for their sound palettes during exploration or extract *sweeteners* or textural components to layer over other sounds (see section 5.1.3). This capability to be able to modify audio signals in novel ways gives creators greater opportunities to create new artifacts.

> ***P4**: "The approach where it is more about creating the individual units of sound rather than the finished product of sound, makes much more sense. It seems at least to be more achievable than what AI seems to be doing in the visual space. Because it doesn't always necessarily understand composition, it gets things roughly in place. What I've seen on people using AI for sound is that it's good to get good approximations, but not necessarily always to do things all the way."*

5.3.3 ***Need for focus on AI for sound design***. Most current research in audio synthesis focuses on music and speech production and very little work exists to model environmental sounds [78]. This feeling was conveyed by P4 during the interview:

> ***P4**: "A lot of the applications you're seeing right now are kind of in the infant stages a lot of the time. From what I've seen so far in sound there haven't been that many great uses of AI so far, at least ones commercially available or available on the market. And a lot of that, I think is because they're taking a more music approach where they're trying to streamline the job of a music producer."*

Further, given the recent surge in text-to-audio models, two designers *(P4, P5)* felt that AI models that needed to be prompted using text would be a barrier for sound design which needs granular, continuous, and *"intimate control" (P2)* to design sounds. Developing controls over AI models where designers can *"leverage their current skills" (P5)* instead of learning newer ways to prompt AI models would be more beneficial for creator use.

6 DISCUSSION

In this study, we sought to investigate how generative AI technologies could support sound design practitioners in their creative work. We found that AI-based CSTs could assist sound designers in their creative process by providing means to iterate over ideas quickly, by generating fantastical and novel-sounding elements, and by reducing the need to manually source individual artifacts via field recording for their creative work. Further, we found that although the unpredictability of controlling the AI-generated artifacts assisted in the serendipitous discovery of new sounds, the exploratory nature and unpredictability in controlling the generation could be a hindrance to task-oriented work. Further, in our study, the sound designers employed various strategies while exploring the design space generated by the AI-based CSTs. These strategies helped them better understand the limitations of the generation capabilities of AI-based tools. Finally, while AI algorithms are usually incentivized to accurately replicate real-world sounds, in contrast, we found that sound designers were more interested in the overall perceptual aspects of the sound than its accuracy. We thus found that AI-based CSTs could easily be integrated as part of a larger creative design process, owned and controlled by the designer. Such CSTs can produce novel sound elements that sound designers can incorporate into their compositions as layers over other sounds or use as individual components for a better cinematic effect than the accuracy in their compositions.

6.1 AI assistance in the practice of sound design

Recently, human-AI interaction researchers have been increasingly interested in understanding how mixed-initiative creative interfaces (MICIs) [29] can be applied in a work setting in different domains of creative work [68, 69, 96]. In our work, we respond to these questions in the context of sound design by proposing a mode of working with generative AI where designers perform exploration and creation using AI-based CSTs. Findings from our exploratory study suggest that such tools can assist in a fast iterative exploration (section 5.1.1) to help sound designers find novel sounds to use in their work. This finding is in line with some recent research on CSTs in the visual domain, in music composition, and in storytelling where algorithmic tools were used predominantly for idea generation [14, 20, 22, 53, 61, 62]. Further, such AI-based tools can generate synthetic surrogates of real-life sensory information (such as, in our case, field recordings (section 5.1.2)) which can constitute realistic and convincing alternatives to this information. Consequently, (sound) designers would be able to save the time and resources they would need to obtain this information in the

first place. This observation could be extended beyond the realm of sound and also include visuals and other sensory modalities.

In [17], researchers note that while the unpredictability (section 5.2) emerging from AI-based tools supports creativity, it could be a hindrance to task-oriented creative work. We further this understanding for sound design (section 5.1.4) and find that sound designers might overcome this limitation by performing exploration (section 5.2.1) as a separately focused task [26], by employing *"reduced listening" (P5)*, to *"build a library" (P4)* of novel sound palettes for use in their projects. The possibility of using CSTs in this way to generate novel individual units of sounds, instead of entire compositions, gives professionals another tool *"in their arsenal" (P4)* and more ownership of their creative process (section 5.3.2).

6.2 Constrained and Unconstrained Randomness

Previously, researchers have investigated the role of constrained and unconstrained randomness in interactive systems on user experience [59, 93]. In [59], using an example of a music-listening interactive system the authors observe that, at times, unconstrained randomness can contribute to rich user experience (such as serendipity). They also note that this positive experience depends upon the size of the audio library, where large-sized libraries can have detrimental effects on the listener experience. In such cases, adding constraints to randomness (by constraining content) gives the users the ability to manipulate or control the affective state of their user experience. We observe this duality of unpredictability and constraint in our study. Our impact sounds 'Hits & Scratches' model was smaller and more constrained in terms of the variety of sounds generated as compared to the 'Environmental Sounds' model which generated sounds from seven classes. Our participants found models with large variances in timbres, such as the environmental sounds model, detrimental to targeted creative exploration. For instance, participant P7 reported: *"The variety of sounds that I got out of the (environmental sounds model) was very extreme. I think that a tool that offers such a broad variety of results is like a two-edged sword."*

Further, our interface-1 was constrained in terms of providing means to explore the AI's latent space using only synthetic sounds, as compared to interface-2 which provided means for unconstrained exploration directly in the latent space of the model. While using our CSTs, P6 reported: *"(Interface-1) was just like playing with an old synthesizer or something. It was quite easy to grab things and just tweak them and see what happened. (With interface-2) none of these settings did anything I was expecting at all."*. Our findings thus indicate that constraints implemented by either smaller models (such as the 'Hits & Scratches' model) or by using synthetic sounds for steering the CSTs assisted designers in better understanding the capabilities of AI (see section 5.2.4) than when using larger models or interface-2.

6.3 Reflections on designing and implementing AI-based tools for sound design

On selecting interactive AI models: While we implemented two CSTs in this study, our aim was not to compare them with each other but to provide our participants with two unique ways of interacting with the underlying AI model. While selecting algorithms for interactivity, we aimed to explore algorithms that worked primarily in a post-hoc fashion (i.e., worked on existing pre-trained GAN models). We found that using methods such as SeFa [79], we could integrate any available pre-trained GAN models from existing marketplaces [31, 46, 103]. Further, using methods such as EBF [51], enabled us not only to use domain-specific controls for exploration but also additionally constrain multi-class large audio models using class-based soft constraints [83]. Using these soft constraints, the designers could target their exploration to a part of the latent space oriented toward that class. We thus found both these methods effective in providing a wide range of options for exploration [81] within our CSTs. Such methodologies for designing interactivity over AI models can be easily extended to other modalities such as images. In light of the recent environmental impact [25] due to the training of large generative AI models, we suggest future CSTs, for all modalities including sound design, could make use of existing pre-trained models by leveraging such post-hoc methods for interactivity.

On visualizing sounds: While designing our interfaces, we visualize the spectrogram of the generated sound because the controls on both interfaces modified the spectromorphology [82] of the sound. Interestingly, through our interviews, we found that these visualizations provided means for the designers to describe their creative goals in spectromorphological terms. For instance, participants used terms such as *"seeing the individual events" (P2)*, *"fade-in is quite long" (P4)*, or *"removing the initial transient and softening it to leave the body and tail" (P4)*, etc. Previously, researchers in the explainable AI (XAI) for arts [12, 13] used latent space visualizations to *explain or debug* their creative goals. We build upon this work and suggest that spectrogram visualizations could provide a great means for designers to communicate their creative goals and to understand the output of AI-based CSTs.

6.4 Ambiguity in interactive user control

On interface-2, we deliberately left the dimensions unlabeled to allow the designers to interpret the dimensions based on their intuition. The ambiguity in the dimensions made the exploration *"more open (P4)"*(section 5.2.2) and different participants came up with different semantic or acoustic explanations for the effect of each dimension on the edited sound (section 5.2.1). Participant P6 reported that Dimension 6 on the interface seemed to semantically change if the source of the sound was *"outside or inside the room"*. Further P1 reported that Dimension 7 and 10 were similar to acoustic high-pass and low-pass filters and P3 commented that Dimension 10 changed the pitch of the sound. By naming the dimensions differently, by using semantic or acoustic labels, the designers were able to use the sound design space in their creative work in a personalized way. Further, with interface-2, participants had to adopt a more varied number of strategies to meaningfully explore the sound design space (section 5.2.1) as compared to interface-1. Therefore, although interface-2 opened up more personalized avenues for the designers to interact with the AI, the ambiguity in the dimensions got in the way of its *agentive flow* [57], a highly engaging state of interacting with an AI-based CST. The ambiguity in the controls in

interface-2 made the designers focus more on the intricacies of the system itself, rather than just focus on their creative output.

7 DESIGN RECOMMENDATIONS FOR HUMAN-AI INTERACTION IN SOUND DESIGN

In this section, we outline five design recommendations for interactive generative AI. We specifically reported some quotes capturing rich insights from our expert practitioners to inspire our readers.

DR1: Design interactivity using intuitive controls. From among our participants, *P2* and *P9* had extensive prior experience designing audio interfaces, synthesizers, and programming desktop foley systems. Their advice on designing a good perceptually relevant set of controls for sound synthesis systems is as follows. They suggest a good control should be:

- **Perceptually monotonic**: If you moved a control forward to change the sound by an X amount, then moving it more in the same direction should do more of X.
- **Perceptually linear**: This principle builds upon monotonic controls. If you moved a control by an X amount in the forward direction, and then you moved it the same amount in the reverse direction, both changes to the sound should be perceptually the same.
- **Perceptually orthogonal**: If you had multiple controls, a change in one control should be independent of others.

These principles are especially important when developing technology specific controls (as on interface-2) as these controls are extracted by an algorithm from the latent space of a generative model. We thus propose future human-AI interaction researchers focus on constraining such algorithms to yield specific changes based on these principles.

DR2: Variety is a two-edged sword. The general trend in large language models or image generation research is to build large overarching generalizable AI models that cater to generating a large variety of images, art, or text. A similar trend is observed in audio, where a large audio model generates music, environmental sounds, as well as speech [6, 60]. Such large audio models can perform well as tools for exploration but are less useful for task-oriented work. This is particularly due to the complexity of the learned latent space. Small changes in the parameter space of such models can lead to large perceptual changes in the generated sounds. Participant P7 termed this variety as a "two-edged sword". We thus propose that future interactive AI applications for sound design focus on giving designers the ability to explore smaller models trained on a more targeted range of sounds. Or provide means to constrain the exploration of large audio models based on class, semantics, or other perceptual aspects of the sound (see section 6.3).

DR3: More cinematic effect than accuracy. In section 5.3.1 and 5.1.3 we showed that our participants valued perceptual aspects of the generated sounds and the AI's ability to generate 'unreal but tangible' sound palettes, more than the accuracy or the origin of the sound. Currently, most audio AI algorithms objectively incentivize the replication of real-world sounds. While real-world sound replications are useful as an alternative to field recording (section 5.1.2), they will have very limited use in being able to generate novel sound palettes. We thus propose that there is value in pursuing a research approach where AI models *"do not replicate real life too well"(P4)* and are able to extract textures and patterns from sounds which current signal processing techniques cannot do. This approach would give artists and creators more creative tools in their arsenal, rather than simply automating the generation of real-world sounds that they can record easily.

DR4: Seeing sounds as an alternative to listening. Previously, Cartwright et al [18] demonstrated that when using visual representations of sounds such as spectrograms, they collected better annotations for sound events than when using audio alone. Visual spectrogram representations of the sounds gave annotators an opportunity to 'glance-and-click' on the sound events while listening which improved the accuracy of the collected annotations. In our study, we make a similar observation. At times, the designers used the spectrograms on the interfaces as a proxy for listening. *"It's very nice to have the spectrogram because this gives you a good forecast. It is a good shortcut to imagine how it will sound like so you can even not listen to it" (P5).* We thus propose that using such visual representations of the sounds can reduce the cognitive load associated with making small edits and stopping to listen to the generated sounds, especially when doing exploratory work.

DR5: Improving the explainability of dimensions. As observed in section 5.2.3 and 5.2.2, although most designers found the ambiguity in dimensions a hindrance to task-oriented work, they observed that giving them the ability to personalize the dimension names would improve the usability of such tools and the explainability of the dimensions (especially with interface-2). *"With the 10-D interface, I found myself wanting to change the label after I explored it so that I could remember what it did for me"* (P2). Further, in our conversations with P6 and P7, we observed that for understanding and learning controls on synthesizer interfaces, designers usually relied on not just the names of the controls, but also their ranges and units of control. For instance, units such as 'dB per octave' are associated with filtering frequencies. P6 observed that on interface-2 all dimensions operated in a range of [-5, +5] with no units, which made it difficult to memorize the function of each control. We thus propose future human-AI interaction research to encompass dimensional controllability for sound models to rescale the ranges and adjust or assign units on controls to fit existing techniques on commercial synthesizer interfaces.

8 LIMITATIONS AND FUTURE WORK

Audio AI research is evolving rapidly with newer innovations in building larger, faster, and better-quality generative audio architectures. While we use StyleGANs [52] to build our CSTs, other alternatives based on AI architectures such as Diffusion [60] are emerging as potential alternatives. Although we have tried to keep our inferences on assessing the potential of AI for sound design free from any technical constraints or usability issues, new modes of interactivity will change how designers perceive and use AI. Therefore, more research will be needed in the future to understand how the practice of sound design evolves along with newer AI models.

We conducted this study with nine professional sound designers from diverse geographic, years of experience, and sound design backgrounds. With this, although we present a rich description of how AI-based CSTs can be used by sound designers in a work setting, given the qualitative nature of our study our findings might not generalize to broader populations. Further, the study was conducted where the participants used the AI CSTs for only a few tasks. Our future work will focus on capturing patterns of usage as well as studying the different parameter exploration strategies in depth in a professional work setting, over longer periods, and in various phases within the sound design project cycle.

9 CONCLUSION

In this paper, we investigated how sound designers can use generative audio AI models in their creative practice. We designed and implemented two interactive audio AI CSTs and invited nine professional sound designers to apply the CSTs in their practice. Through semi-structured interviews, we gathered insights on how to situate AI-based tools in the sound design process, the sound designer's ways of working with unpredictability and ambiguity in AI, and their expectations of generative AI-based tools. Further, we reported five design recommendations for future interactive AI-based creative support tools for sound design. Through this work, we hope to bring focus to this area of interactive audio AI and explore opportunities to improve AI assistance in the practice of sound design.

ACKNOWLEDGMENTS

We sincerely thank all the sound design experts involved in our research for their time and valuable insights on AI-based Creative Support Tools during our discussion. We would also like to thank Hannah Qiao for her creative design of the introductory video for this research.

REFERENCES

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* (2023).

[2] Open AI. 2023. Dall.E 2. https://openai.com/dall-e-2 [Accessed: 29 August 2023].

[3] Open AI. 2023. Introducing ChatGPT. https://openai.com/blog/chatgpt [Accessed: 29 August 2023].

[4] Aimi.Fm. 2023. Aimi.Fm. https://www.aimi.fm/ [Accessed: 15 November 2023].

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[6] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 2523–2533. https://doi.org/10.1109/TASLP.2023.3288409

[7] Renaud Bougueng Tchemeube, Jeffrey John Ens, and Philippe Pasquier. 2022. Calliope: A Co-creative Interface for Multi-Track Music Generation. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 608–611. https://doi.org/10.1145/3527927.3535200

[8] brain.fm. 2023. brain.fm. https://www.brain.fm/ [Accessed: 15 November 2023].

[9] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[10] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (2021), 328–352. https://doi.org/10.1080/14780887.2020.1769238

[11] Virginia Braun and Victoria Clarke. 2021. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health* 13, 2 (2021), 201–216. https://doi.org/10.1080/2159676X.2019.1704846

[12] Nick Bryan-Kinns, Berker Banar, Corey Ford, Courtney N. Reed, Yixiao Zhang, Simon Colton, and Jack Armitage. 2021. Exploring XAI for the Arts: Explaining Latent Space in Generative Music, In eXplainable AI approaches for debugging and diagnosis. *XAI 4 Debugging Workshop at NEURIPS 2021*. https://openreview.net/forum?id=GLhY_0xMLZr

[13] Nick Bryan-Kinns, Corey Ford, Alan Chamberlain, Steven David Benford, Helen Kennedy, Zijin Li, Wu Qiong, Gus G. Xia, and Jeba Rezwana. 2023. Explainable AI for the Arts: XAIxArts. In *Proceedings of the 15th Conference on Creativity and Cognition* (Virtual Event, USA) *(C&C '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3591196.3593517

[14] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *HAI-GEN+ user2agent@ IUI*.

[15] Linda Candy. 2006. Practice based research: A guide. *CCS report* 1, 2 (2006), 1–19.

[16] Linda Candy, Shigeki Amitani, and Zafer Bilda. 2006. Practice-led strategies for interactive art research. *CoDesign* 2, 4 (2006), 209–223. https://doi.org/10.1080/15710880601007994

[17] Baptiste Caramiaux and Sarah Fdili Alaoui. 2022. "Explorers of Unknown Planets": Practices and Politics of Artificial Intelligence in Visual Arts. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 477 (nov 2022), 24 pages. https://doi.org/10.1145/3555578

[18] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. 2017. Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 29 (dec 2017), 21 pages. https://doi.org/10.1145/3134664

[19] Michel Chion. 2019. *Audio-vision: sound on screen*. Columbia University Press.

[20] Li-Yuan Chiou, Peng-Kai Hung, Rung-Huei Liang, and Chun-Teng Wang. 2023. Designing with AI: An Exploration of Co-Ideation with Image Generators. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1941–1954. https://doi.org/10.1145/3563657.3596001

[21] Keunwoo Choi, Jaekwon Im, Laurie Heller, Brian McFee, Keisuke Imoto, Yuki Okamoto, Mathieu Lagrange, and Shinosuke Takamichi. 2023. Foley sound synthesis at the dcase 2023 challenge. *arXiv preprint arXiv:2304.12521* (2023).

[22] John Joon Young Chung, Shiqing He, and Eytan Adar. 2021. The Intersection of Users, Roles, Interactions, and Technologies in Creativity Support Tools. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1817–1833. https://doi.org/10.1145/3461778.3462050

[23] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[24] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and Controllable Music Generation. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=jtiQ26sCJi

[25] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

[26] Mihaly Csikszentmihalyi. 2014. *Toward a Psychology of Optimal Experience*. Springer Netherlands, Dordrecht, 209–226. https://doi.org/10.1007/978-94-017-9088-8_14

[27] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko. 2016. Empirically Studying Participatory Sense-Making in Abstract Drawing with a Co-Creative Cognitive Agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 196–207. https://doi.org/10.1145/2856767.2856795

[28] Brecht De Man, Ryan Stables, and Joshua D Reiss. 2019. *Intelligent Music Production*. Routledge.

[29] Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, Marco Gillies, Jeremy Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. 2017. Mixed-Initiative Creative Interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA,

628–635. https://doi.org/10.1145/3027063.3027072
[30] Endel. 2023. Endel. https://endel.io/ [Accessed: 15 November 2023].
[31] Hugging Face. 2023. Hugging Face. https://huggingface.co/ [Accessed: 15 November 2023].
[32] Rebecca Fiebrink and Baptiste Caramiaux. 2016. The machine learning algorithm as creative musical tool. *arXiv preprint arXiv:1611.00379* (2016).
[33] Rebecca Fiebrink and Perry R Cook. 2010. The Wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, Vol. 3. Citeseer, 2–1.
[34] Corey Ford and Nick Bryan-Kinns. 2022. Speculating on Reflection and People's Music Co-Creation with AI. *Workshop on Generative AI and HCI at the CHI Conference on Human Factors in Computing Systems 2022* (2022).
[35] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the Landscape of Creativity Support Tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3290605.3300619
[36] Emma Frid, Celso Gomes, and Zeyu Jin. 2020. Music Creation by Example. 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, 1–13. https://doi.org/10.1145/3313831.3376514
[37] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/642611.642653
[38] Frederic Gmeiner, Humphrey Yang, Lining Yao, Kenneth Holstein, and Nikolas Martelaro. 2023. Exploring Challenges and Opportunities to Support Designers in Learning to Co-create with AI-based Manufacturing Design Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 226, 20 pages. https://doi.org/10.1145/3544548.3580999
[39] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
[40] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA, USA. http://www.deeplearningbook.org.
[41] Chitralekha Gupta, Purnima Kamath, Yize Wei, Zhuoyao Li, Suranga Nanayakkara, and Lonce Wyse. 2023. Towards Controllable Audio Texture Morphing. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096328
[42] Chitralekha Gupta, Purnima Kamath, and Lonce Wyse. 2021. Signal representations for synthesizing audio textures with generative adversarial networks. In *Proceedings of the 18th Sound and Music Computing Conference*, Simone Spagnol Davide Andrea Mauro and Andrea Valle (Eds.). Sound and Music Computing Network, Axea sas/SMC Network. https://doi.org/10.5281/zenodo.5113511
[43] Aaron Hertzmann. 2018. Can computers create art?. In *Arts*, Vol. 7. MDPI, 18.
[44] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6840–6851. https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
[45] 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. *arXiv preprint arXiv:2010.05388* (2020).
[46] Pytorch Hub. 2023. Pytorch Hub. https://pytorch.org/hub/ [Accessed: 15 November 2023].
[47] Daniel Hug and Nicolas Misdariis. 2011. Towards a conceptual framework to integrate designerly and scientific sound design methods. In *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound* (Coimbra, Portugal) *(AM '11)*. Association for Computing Machinery, New York, NY, USA, 23–30. https://doi.org/10.1145/2095667.2095671
[48] Muhammad Huzaifah and Lonce Wyse. 2021. *Deep Generative Models for Musical Audio Synthesis*. Springer International Publishing, Cham, 639–678. https://doi.org/10.1007/978-3-030-72116-9_22
[49] Robert Jack, Jacob Harrison, and Andrew McPherson. 2020. Digital Musical Instruments as Research Products. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 446–451. https://doi.org/10.5281/zenodo.4813465
[50] Youngseung Jeon, Seungwan Jin, Patrick C. Shih, and Kyungsik Han. 2021. FashionQ: An AI-Driven Creativity Support Tool for Facilitating Ideation in Fashion Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 576, 18 pages. https://doi.org/10.1145/3411764.3445093
[51] Purnima Kamath, Chitralekha Gupta, Lonce Wyse, and Suranga Nanayakkara. 2023. Example-Based Framework for Perceptually Guided Audio Texture Generation. arXiv:2308.11859 [eess.AS]
[52] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813
[53] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale Text-to-Image Generation Models for Visual Artists' Creative Works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 919–933. https://doi.org/10.1145/3581641.3584078
[54] Janin Koch, Nicolas Taffin, Andrés Lucero, and Wendy E. Mackay. 2020. SemanticCollage: Enriching Digital Mood Board Design with Semantic Labels. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 407–418. https://doi.org/10.1145/3357236.3395494
[55] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=a-xFK8Ymz5J
[56] Leigh Landy. 2007. *Understanding the art of sound organization*. Mit Press.
[57] Tomas Lawton, Kazjon Grace, and Francisco J Ibarrola. 2023. When is a Tool a Tool? User Perceptions of System Agency in Human–AI Co-Creative Drawing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1978–1996. https://doi.org/10.1145/3563657.3595977
[58] Sara Lenzi. 2021. The design of data sonification. Design processes, protocols and tools grounded in anomaly detection. (2021).
[59] Tuck Wah Leong, Frank Vetere, and Steve Howard. 2006. Randomness as a Resource for Design. In *Proceedings of the 6th Conference on Designing Interactive Systems* (University Park, PA, USA) *(DIS '06)*. Association for Computing Machinery, New York, NY, USA, 132–139. https://doi.org/10.1145/1142405.1142428
[60] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 21450–21474. https://proceedings.mlr.press/v202/liu23f.html
[61] Ryan Louie, Andy Coenen, Cheng Zhi 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, 1–13. https://doi.org/10.1145/3313831.3376739
[62] Ryan Louie, Jesse Engel, and Cheng-Zhi Anna Huang. 2022. Expressive Communication: Evaluating Developments in Generative Models and Steering Interfaces for Music Creation. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 405–417. https://doi.org/10.1145/3490099.3511159
[63] Yuwen Lu, Chengzhi Zhang, Iris Zhang, and Toby Jia-Jun Li. 2022. Bridging the Gap Between UX Practitioners' Work Practices and AI-Enabled Design Support Tools. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 268, 7 pages. https://doi.org/10.1145/3491101.3519809
[64] Todd Lubart. 2005. How can computers be partners in the creative process: Classification and commentary on the Special Issue. *International Journal of Human-Computer Studies* 63, 4 (2005), 365–369. https://doi.org/10.1016/j.ijhcs.2005.04.002 Computer support for creativity.
[65] Andrés Marafioti, Nathanaël Perraudin, Nicki Holighaus, and Piotr Majdak. 2019. Adversarial Generation of Time-Frequency Features with application in audio synthesis. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 4352–4362. https://proceedings.mlr.press/v97/marafioti19a.html
[66] Fabio Morreale and Antonella De Angeli. 2016. Collaborating with an autonomous agent to generate affective music. *Computers in Entertainment (CIE)* 14, 3 (2016), 1–21.
[67] Mohammad Amin Mozaffari, Xinyuan Zhang, Jinghui Cheng, and Jin L.C. Guo. 2022. GANSpiration: Balancing Targeted and Serendipitous Inspiration in User Interface Design with Style-Based Generative Adversarial Network. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 537, 15 pages. https://doi.org/10.1145/3491102.3517511

[68] Michael Muller, Lydia B Chilton, Anna Kantosalo, Q. Vera Liao, Mary Lou Maher, Charles Patrick Martin, and Greg Walsh. 2023. GenAICHI 2023: Generative AI and HCI at CHI 2023. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 350, 7 pages. https://doi.org/10.1145/3544549.3573794

[69] Michael Muller, Lydia B Chilton, Anna Kantosalo, Charles Patrick Martin, and Greg Walsh. 2022. GenAICHI: Generative AI and HCI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 110, 7 pages. https://doi.org/10.1145/3491101.3503719

[70] Leo Murray. 2019. *Sound design theory and practice: Working with sound.* Routledge.

[71] 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)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174223

[72] Sangshin Oh, Minsung Kang, Hyeongi Moon, Keunwoo Choi, and Ben Sangbae Chon. 2023. A Demand-Driven Perspective on Generative Audio AI. arXiv:2307.04292 [eess.AS]

[73] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. 2016. Visually Indicated Sounds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 2405–2413. https://doi.org/10.1109/CVPR.2016.264

[74] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[75] Zdeněk Průša, Peter Balazs, and Peter Lempel Søndergaard. 2017. A Noniterative Method for Reconstruction of Phase From STFT Magnitude. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 5 (2017), 1154–1164. https://doi.org/10.1109/TASLP.2017.2678166

[76] Jeba Rezwana and Mary Lou Maher. 2023. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 67 (sep 2023), 28 pages. https://doi.org/10.1145/3519026

[77] Hugo Scurto, Bavo Van Kerrebroeck, Baptiste Caramiaux, and Frédéric Bevilacqua. 2021. Designing Deep Reinforcement Learning for Human Parameter Exploration. *ACM Trans. Comput.-Hum. Interact.* 28, 1, Article 1 (jan 2021), 35 pages. https://doi.org/10.1145/3414472

[78] Garima Sharma, Karthikeyan Umapathy, and Sridhar Krishnan. 2022. Trends in Audio Texture Analysis, Synthesis, and Applications. *Journal of the Audio Engineering Society* 70, 3 (2022), 108–127. https://doi.org/10.17743/jaes.2021.0060

[79] Yujun Shen and Bolei Zhou. 2021. Closed-Form Factorization of Latent Semantics in GANs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 1532–1540. https://doi.org/10.1109/CVPR46437.2021.00158

[80] Ben Shneiderman. 2007. Creativity Support Tools: Accelerating Discovery and Innovation. *Commun. ACM* 50, 12 (dec 2007), 20–32. https://doi.org/10.1145/1323688.1323689

[81] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.

[82] Denis Smalley. 1997. Spectromorphology: explaining sound-shapes. *Organised sound* 2, 2 (1997), 107–126.

[83] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4080–4090.

[84] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=St1giarCHLP

[85] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. 2021. Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 1415–1428. https://proceedings.neurips.cc/paper_files/paper/2021/file/0a9fdbb17feb6ccb7ec405cfb85222c4-Paper.pdf

[86] Angie Spoto, Natalia Oleynik, Sebastian Deterding, and Jon Hook. 2017. Library of Mixed-Initiative Creative Interfaces.

[87] Christian J. Steinmetz and Joshua Reiss. 2021. pyloudnorm: A simple yet flexible loudness meter in Python. In *Audio Engineering Society Convention 150*. http://www.aes.org/e-lib/browse.cfm?elib=21076

[88] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 582, 11 pages. https://doi.org/10.1145/3411764.3445219

[89] Patrick Susini, Olivier Houix, and Nicolas Misdariis. 2014. Sound design: an applied, experimental framework to study the perception of everyday sounds. *The New Soundtrack* 4, 2 (2014), 103–121.

[90] Koray Tahiroğlu, Miranda Kastemaa, and Oskar Koli. 2020. Al-terity: Non-Rigid Musical Instrument with Artificial Intelligence Applied to Real-Time Audio Synthesis. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 337–342. https://doi.org/10.5281/zenodo.4813402

[91] Akito Van Troyer and Rebecca Kleinberger. 2019. From Mondrian to Modular Synth: Rendering NIME using Generative Adversarial Networks. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Marcelo Queiroz and Anna Xambó Sedó (Eds.). UFRGS, Porto Alegre, Brazil, 272–277. https://doi.org/10.5281/zenodo.3672956

[92] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *CoRR* abs/1609.03499 (2016). arXiv:1609.03499 http://arxiv.org/abs/1609.03499

[93] Jhim Kiel M. Verame, Enrico Costanza, Joel Fischer, Andy Crabtree, Sarvapali D. Ramchurn, Tom Rodden, and Nicholas R. Jennings. 2018. Learning from the Veg Box: Designing Unpredictability in Agency Delegation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174021

[94] Prateek Verma and Chris Chafe. 2021. A generative model for raw audio using transformer architectures. In *2021 24th International Conference on Digital Audio Effects (DAFx)*. IEEE, 230–237. https://www.dafx.de/paper-archive/2021/proceedings/papers/DAFx20in21_paper_40.pdf

[95] Graham Wallas. 1926. *The art of thought.* Vol. 10. Harcourt, Brace.

[96] Justin D. Weisz, Michael J. Muller, Jessica He, and Stephanie Houde. 2023. Toward General Design Principles for Generative AI Applications 130-144. In *IUI Workshops*. https://api.semanticscholar.org/CorpusID:255825625

[97] Lilian Weng. 2021. What are diffusion models? *lilianweng.github.io* (Jul 2021). https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

[98] Anna Wiener. 2022. The Weird, Analog Delights of Foley Sound Effects. https://www.newyorker.com/magazine/2022/07/04/the-weird-analog-delights-of-foley-sound-effects

[99] Lonce Wyse. 2018. Real-valued parametric conditioning of an RNN for interactive sound synthesis. *arXiv preprint arXiv:1805.10808* (2018).

[100] Lonce Wyse, Purnima Kamath, and Chitralekha Gupta. 2022. Sound Model Factory: An Integrated System Architecture for Generative Audio Modelling. In *Artificial Intelligence in Music, Sound, Art and Design*, Tiago Martins, Nereida Rodríguez-Fernández, and Sérgio M. Rebelo (Eds.). Springer International Publishing, Cham, 308–322.

[101] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023. Diffsound: Discrete Diffusion Model for Text-to-Sound Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 1720–1733. https://doi.org/10.1109/TASLP.2023.3268730

[102] 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* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 43–47. https://doi.org/10.1145/3397481.3450663

[103] Model Zoo. 2023. Model Zoo. https://modelzoo.co/ [Accessed: 15 November 2023].

A APPENDIX: SUPPLEMENTARY MATERIALS

A.1 Semi-structured Interview Questions

As discussed in section 4.2, our interview consisted of three parts:

- Participant's background and experience: Through these questions, we focused on capturing the participant's experience with sound design
  - Can you describe some of the projects that you typically work with?

  - Can you describe with an example some of the typical tasks you perform while designing sounds?
  - What is the most annoying part of your design process?
- Their expectations of generative AI: These questions captured the participant's outlook and past experience with generative AI.
  - What do you feel about AI?
  - What were your expectations from this AI-based sound design tool?
  - Can you describe an ideal AI-based tool for sound design?
- Based on the creative task: Through these questions, we capture the participant's experience and feedback using the AI-based sound design tools in this study. For this part of the interview, we use the screen recordings as discussion prompts.
  - Can you explain what you wanted to do in the open-ended task and how did you go about achieving it?
  - Did the outcomes from your tasks match your expectations?
  - How do you think such AI-assisted sound design tools fit into your design process?
  - What did you find most frustrating to do?
  - What do you want AI-assisted sound design tools to do more?

A.2 AI-based CST Architecture Details

The two interfaces in this study, interface-1 which uses domain-specific controls, and interface-2 which uses technology-specific controls, use the same underlying trained StyleGAN. The StyleGAN architecture is shown in Figure 4 (a), where G_s is the generator (synthesis network), G_m is the mapping network, and D is the discriminator. E is a GAN inversion network adapted from [51] for interface-1. A StyleGAN2's generator can be modeled as a function $G(.)$ that maps a latent space $\mathcal{Z}$, where $\mathbf{z} \in \mathbb{R}^{\delta_z}$, to the higher dimensional spectrogram space $\mathbf{S} \in \mathbb{R}^{f \times t}$, such that $\mathbf{S} = G(\mathbf{z})$. Here δ_z is the dimensionality of the $\mathcal{Z}$ space and f, t are the number of frequency channels and time frames of the generated spectrogram respectively. StyleGANs further learn an intermediate representation $\mathcal{W}$, where $\mathbf{w} \in \mathbb{R}^{\delta_w}$, between that of $\mathcal{Z}$ and $\mathcal{S}$ via a mapping network $G_m(.)$. This intermediate latent space further disentangles factors of variation as compared to the latent $\mathcal{Z}$ space [52]. Further, a synthesis network $G_s(.)$ maps the $\mathbf{w}$ vector to a spectrogram $\mathbf{S}$. Note that in this paper, whenever we refer to the term "latent space", we mean the $\mathcal{W}$-space generated by the mapping network G_m.

We set $\mathcal{Z}$ and $\mathcal{W}$ space dimensions δ_z and δ_w both to 128 and use 4 mapping layers in the Generator for all our experiments. Further, we use the log-magnitude spectrogram representations generated using a Gabor transform [65](n_frames= 256, stft_channels= 512, hop_size= 128), a Short-Time Fourier Transform (STFT) with a Gaussian window, to train the StyleGAN2, and the Phase Gradient Heap Integration (PGHI) [75] for high-fidelity spectrogram inversion of textures to audio [42]. All sounds generated using both interfaces were normalized to -14dB for loudness using pyloudnorm [87]. The codebase for the interfaces, StyleGAN, and Encoder is on GitHub as follows:

- Both interfaces in this study: https://github.com/augmented-human-lab/audio-design-toolkit
- StyleGAN: https://github.com/pkamath2/audio-stylegan2
- GAN Encoder: https://github.com/pkamath2/audio-latent-composition

A Google Colaboratory version of our interactive Creative Support Tools can be found here: https://pkamath2.github.io/chi2024-resources/

A.2.1 Interface-1. Apart from StyleGAN, interface-1 is powered by two additional components: (1) a GAN Encoder or inversion framework, and (2) a synthetic sound generator. The code for both is adapted from [51]. While GANs map the latent space to real-world sounds, GAN Encoders learn the inverse, i.e. they map real-world sounds to latent space embeddings. This technique is especially useful when we want to "query" or "search" sounds within the latent space using approximations (or synthetic sounds). The synthetic sounds are generated by passing Gaussian noise $\mathcal{N}(0, I)$ through band-pass and fade filters. The parameters to generate the synthetic sounds are actualized as sliders on the interface. Figure 4 (b) shows a conceptual diagram for the components behind interface-1. The synthetic sounds are encoded into the latent space to derive their $\mathcal{W}$-embeddings. These embeddings are then passed through the generator to generate realistic AI-generated sounds matching the synthetic sounds.

A.2.2 Interface-2. We utilize the semantic factorization algorithm (SeFa) [79] to derive technology-specific controls from the latent space of the StyleGAN in this study. The SeFa method is a closed-form unsupervised method for latent semantic discovery. It decomposes the pre-trained weights of G_m of the StyleGAN using eigendecomposition to find vectors for controllability. The SeFa algorithm returns δ_w (128 in our case) dimension vectors and their corresponding singular values. We fetch the top 10 vectors (vectors with the highest singular values) and display them on the interface. The vectors are actualized as sliders on the interface for users to interact and perform edits directly in the latent space of the GAN.

Both interfaces were developed using Streamlit and ReactJs. Streamlit is a Python library that enables frontend applications to connect to Python-based machine learning models easily. The ReactJs-based frontend communicates with the Python backend using Websockets.

A.3 Acoustic Parameters on Interface-1

The list of acoustic parameters on the interface-1 are:

(1) Impulse width: Parameter value decides how long the impact sound 'rings' or lasts along the time axis.
(2) Rate: Controls the number of impact events in the sound along the time axis.
(3) Frequency band: Frequency range of the bandpass filters. Controls the brightness of the sound. Higher frequencies sound brighter, such as impact sound on a hard metal surface. Lower frequency ranges sound duller, such as impact sounds on soft materials such as a cushion or a sofa.
(4) Filter order: Determines the frequency roll-off. Used in conjunction with the frequency band. Higher filter orders have a steeper roll-off and transition between the frequency bands.

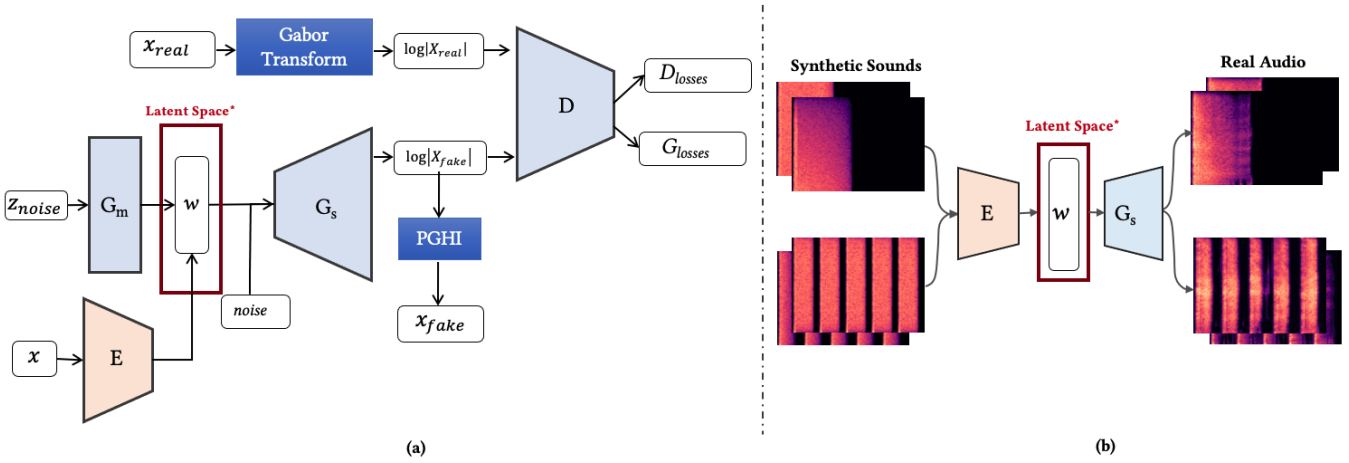

Figure 4: Architectural components driving the audio AI interfaces used in the study

(5) Fade In: Controls how the sound transitions from zero to full strength.
(6) Fade Out: Controls how the sound transitions from full strength to zero.

A.4 Attribution for icons and images

Most images in this paper were created by the authors using a combination of various drawing tools. Some visual icons were sourced from the following websites:

- In Figure 1: the sound designer icon is sourced from Flaticon.com; the domain-specific controls icon is sourced from a "slider" icon by Inggit Jaya from thenounproject.com.
- In Figure 2: the sound designer icon is sourced from Flaticon.com;