AI READS FLUXUS: TEXT SCORES, LANGUAGE GAMES, AND THE ART OF BEING LITERAL

Nicola Privato
Intelligent Instruments Lab
University of Iceland
Reykjavík, Iceland
nprivato@hi.is

Thor Magnusson
Intelligent Instruments Lab
University of Iceland
Reykjavík, Iceland
thormagnusson@hi.is

ABSTRACT

In artistic practices combining the use of text scores with generative AI, two distinct approaches seem to emerge. On the one hand, are works in which LLMs are used to generate scores to be enacted by humans; on the other, are works in which human-generated scores are used to live-prompt text-to-audio models. In this paper, we investigate the latter approach, expanding existing practices through a pipeline for neural synthesis-based manipulation and layering of sonic cues generated live from text scores. We introduce Mouja+, a performance piece in which text scores from Fluxus artists are prompted to a text-to-audio model (Stable Audio Open) and dynamically processed through timbre transfer and spatialisation techniques. Based on the experience of composing and performing Mouja+, we discuss how the way we understand and use language with text-to-audio models needs to change, and identify a series of compositional strategies to address this challenge. We also show how transparency, open access, and searchability in the training data favours an intentional and informed engagement with the source material that greatly enhances the compositional process, an aspect that is also crucial to our everyday engagement with AI systems.

1. INTRODUCTION

Over the past two decades, Machine Learning (ML) technologies have undergone significant advancements. The availability of large datasets, the scaling in computational resources, and the emergence of novel architectures and optimization techniques have extended the application of such systems in a plethora of domains, including the arts.

The growing integration of different types of AI tools in artistic practices is particularly evident in music composition and performance. Symbolic systems are being increasingly applied in the algorithmic generation of musical scores and in the design of conversational tools trained on small and large MIDI datasets [1, 2]; Neural Audio Synthesis (NAS) engines are being thoroughly investigated in NIME [3, 4, 5]; large language models, either alone or combined with diffusion models in text-to-audio systems, are increasingly entering composers' creative processes [6, 3].

In this paper, we examine how artists are re-imagining the established practice of composing and performing text scores through AI-based Natural Language Processing (NLP) techniques. Rather than focusing on the relatively more common application of LLMs to generate text scores, we look at the emerging practice of composing human-authored text scores instructing text-to-audio models' live generation of sound [7]. Our practice-based investigation centres on Mouja+, a performance piece in which text scores from Fluxus artists are adapted for and prompted live to an open-source text-to-audio model, whose outputs are layered into an overarching narrative.

We offer three contributions that respond to research objectives emerging from and informing the artistic practice. First, we address text-to-audio models' limitations in real-time sound manipulation through a performative pipeline that combines asynchronous text-to-audio generation with real-time timbre transfer and spatial audio control. Second, we examine what it means to compose text scores for generative models, discussing how our understanding and use of language needs to adapt, and offer a series of practical strategies. Third, we demonstrate how the increased understanding of the model that derives from the public availability and searchability of the data greatly facilitates the design of text scores. In the last part of the paper, we extend this insight beyond artistic practice, to reflect on how the access to and searchability of the training data might foster an informed and effective interaction with AI systems.

2. BACKGROUND

In Western musical tradition, the role of textual cues has been notoriously that of providing expressive and stylistic clarifications complementing standard notation. However, already at the dawn of modernism, Erik Satie began extending their use to introduce elements of indeterminacy. Cues such as 'de clairvoyance,' or 'ouvrez la tête,' introduced a playful interpretative openness, whimsically engaging the performer through a series of interpretative riddles. This openness was further expanded by the early avant-garde movements, from futurist Luigi Russolo, who championed an art of noises that incorporates the 'sounds of language' [8] to the Dada movement's experimentation

with poetry as sound [9].

Cage represents a turning point in the evolution of text scores, both as a composer and as a mentor in his course at the New School for Social Research throughout the Sixties. His class on composition may be considered the cradle of what would have become the diverse cluster of Fluxus, a collective whose action notation and provocative event scores, systematically challenging assumptions and investigating the limits of performativity and meaning, gained widespread influence in the second half of the 20th century [10].

While Fluxus works are often described as *event scores*, Pauline Oliveros' pieces using natural language are typically categorized under the broader definition of text scores. This difference reflects perhaps Oliveros' unique approach, where the core of the performative experience relies, rather than on action, on the stillness of listening with one's ears and body. Oliveros describes her scores as 'algorithmic improvisation or compositions,' and 'recipes that allow musicians to create music without reading notes' [11]. Text serves here as the means to enter the *deep listening* experience, to investigate the network where information is produced and acted upon. In Oliveros' work, the score's openness foregrounds a relational dimension, encompassing the human and non-human (whether natural or technical) actors embraced by the deep listening practice.

For the purposes of this paper, Oliveros' algorithmic compositions may be considered something different altogether from the algorithms explored, for instance, by Xenakis or Hiller [12], or even from live coding approaches based on formal and domain-specific languages. Indeed, Oliveros' algorithms are based on natural language, a system of communication long considered uniquely accessible to human cognition. Only recently, and especially with advances in machine learning and Natural Language Processing (NLP) techniques, has this mode of communication become accessible to computational systems.

2.1 AI generated Text Scores

Despite recent improvements and exponential diffusion of Large Language Models (LLMs) such as ChatGPT, DeepSeek, Gemini and LLaMa, and their application across disciplines, their use for the generation text scores, whether in music performance and composition or in theatre and dance, remains relatively limited.

A relevant example of artistic practice featuring AI-generated scores comes from Jennifer Walshe. Her corpus of works focusing on voice, identity, hybridity, and deeply shaped by digital culture, includes a massive dataset of approximately 3,500 text scores collected from 2017 to 2021, to be used as training data for large language models [6]. This dataset contains text scores from Oliveros, George Brecht, Mieko Shiomi, and others, but is also open to the contribution of any artist willing to share their work. A first application of this material for the training of an LLM comes from Walshe's collaboration with PRiSM team for the 2021 Darmstadt summer course [13].

Synthetic Erudition Assist Lattice, by the distributed collective Seals, takes instead a more nuanced approach to the use of AI-generated text [3]. The piece features readings from GPT-generated dialogues, but in a political twist these are inserted into a much broader palimpsest, including theremins and the glitches of the remote communication platform used for the performance.

Expanding from music performance into theatre and dance, an example of what might be viewed as AI-generated action scores is that of Orange Grove Dance's A&I [1] theatrical play, featuring *Laura,* an LLM generating real-time textual instructions for a group of performers, de facto choreographing the whole piece and turning it into a destabilizing and surreal experience.

The relatively limited adoption of AI for text score generation is perhaps symptomatic of a certain scepticism about LLMs' creative potential, at least as far as the structuring of a piece is concerned. Even, one might argue, of a certain hesitancy among artists to give up on their agency upon the macro-formal aspects of a piece to follow the machine's lead. Crucially enough, in Walshe, Seals, and Orange Grove Dance, AI functions indeed primarily as a source of indeterminacy, of entropy, in Eco's use of the term [14]. Its value seems to reside precisely in how it misses (and messes) the point, in how it flavours the source material [6] through gaps and glitches [3]. In other words, in its ability to fail upward.

2.2 Text Scores and Text-to-audio

The projects described in Section 2.1 focus specifically on the algorithmic generation of textual instructions, with human performers enacting them. However, the recent emergence and evolution of Neural Audio Synthesis (NAS) techniques, where sounds are synthesized from scratch by generative algorithms, paired with LLMs in text-to-audio models, enables a different approach to combining text scores and AI. Tools such as Udio, Suno AI, and Stable Audio [15], capable of synthesizing coherent musical material from natural language descriptors, allow for the machinic enactment of human-composed text scores.

The first (and to our knowledge only) example of a performative use of text-to-audio NAS in a live setting is represented by Dadabots' *prompt jockeying* live sets [7]. In prompt jockeying, algorithms generate music on stage based on live-prompted textual cues. According to Dadabots, 'if a DJ must be fluent in their Rekordbox collection, a PJ must be fluent in their PyTorch model' [16]. Such fluidity presupposes high programming skills, but also requires many hours of experience in prompting on a particular model. This is due to how LLMs understand and interpret natural language, challenging assumptions about human-to-human communication through natural language, and requiring composers to adapt to and account for the models' interpretative limitations.

Prompt jockeying is based on models that generate audio files asynchronously. This entails (i) a non-negligible time lag between the entering of a prompt and the playback of the generated file (dependent on GPU, sampling rate and length of the file, but usually in the order of tens

[1] https://www.orangegrovedance.com/a-i

of seconds) and (ii) no real-time manipulation over the output through NAS. As an alternative to DJing, prompt jockeying requires minimal human intervention. AI-generated files can be easily overlapped, layered, stretched, and filtered through digital signal processing (DSP) techniques. On the contrary, a performative practice focusing on the real-time intervention through NAS itself inevitably suffers from these technical limitations.

Mouja+, the piece we discuss in the next chapter, addresses this challenge by introducing a performative pipeline for the real-time generation and manipulation of sound from text using NAS, thus extending the use of NAS in combination with text scores from prompt jockeying into live performance. Furthermore, it offers a practical case-study for a discussion on what it means to compose text scores for a text-to-audio model, how our use of language needs to change as we address such systems, and how these strategies might contribute to the broader discourse around AI and data accessibility.

3. MOUJA+

Mouja+ is a piece composed and performed by the first author for the SWRL festival in Riga, Latvia. [2] The festival, focusing on music AI and quantum computing techniques, was held in November 2024 in a 36+2 channels Ambisonics dome. For this occasion, the piece was designed around an existing performance set titled Mouja [17], which employs custom interfaces for real-time NAS performance [18], while incorporating a text-to-audio model prompted with adapted text scores from the Fluxus Performance Workbook [19].

The choice of Fluxus is deliberate. This collective's work systematically investigates the limits of performativity while challenging fundamental assumptions about meaning-making in the arts. Their irreverent, iconoclastic approach, combined with emphasis on embodiment and performativity, provides an ideal foundation for exploring how AI models might push the boundaries of language, communication, and performance.

The following sections outline the software pipeline supporting this investigation.

3.1 Pipeline

Mouja+'s pipeline is based on a combination of two different NAS models, one featuring real-time timbre transfer and latent manipulation, the other affording the asynchronous generation of audio files from textual prompts.

3.1.1 RAVE

RAVE, the first of the two models, is a Variational Autoencoder for real-time timbre transfer [20]. It trains on relatively small amounts of audio (from one to many hours of recordings), learning a compressed, multidimensional representation of the most meaningful timbral parameters

[2] This paper presents outcomes from theoretical reflections of both authors upon the practice of the first author. Where I is used, it refers to the practice-based research from the first author

Figure 1. The Ambisonics Dome at SWRL Festival, during Mouja+ Performance.

of the sound material, in-between an encoding and a decoding function, described as *latent space*. Besides offering high quality, real-time timbre transfer capabilities, RAVE supports direct access and manipulation of its latent space through control signals, alone or combined with timbre transfer functions.

Along with the high quality of the generated audio, one of the most appreciated features of RAVE is its ability to process sounds in real-time. Artists often use RAVE as the new synthesizer in their rig or as the engine for a novel musical interface [21, 5]. Still, compared to traditional synthesis methods, RAVE informs a different thinking about controlling and manipulating sound, one in which sound parameters and latent dimensions are deeply entangled. In previous works, we explored three types of entanglement, showing how these influence the way we play, and proposed a series of tailored performative and compositional approaches, including the development of dedicated interfaces [22]. Even though this paper does not focus on the interfaces themselves, it is worth noticing that in Mouja+ RAVE's latents are manipulated through Stacco, one of the interfaces emerging from our work at the Intelligent Instruments Lab (IIL).

3.1.2 Stable Audio

Stable Audio Open [15], the second NAS model used in Mouja+, is a text-to-audio diffusion model trained on a dataset collected from the Freesound archive [23]. Diffusion models generate sound files by progressively reversing a noise-adding process used for the training, and are often conditioned on text vectors generated by LLMs. The user interacts with such systems by entering positive and negative prompts using natural language, and by defining parameters such as temperature and steps, corresponding to the randomness and the number of denoising iterations in the generation process.

Crucially, whereas real-time autoencoders such as RAVE continuously output small blocks of sounds (typically 2048 samples per block), thus allowing one to dynamically intervene in the timbre transfer process through the manipulation of the latent space, text-to-audio diffusion models asynchronously generate one audio file per prompt. This asynchronous process limits real-time NAS manipulation

Figure 2. GUI for text prompting and timbre transfer.

but enables macro-formal compositional control through natural language instructions.

3.1.3 RAVE and Stable Audio

Stable Audio and RAVE serve two different and yet potentially complementary purposes. On the one hand, Stable Audio invites the composer to work on structure and form through textual instructions, but offers no real-time control over the timbral qualities of the sound; on the other hand, RAVE's small block size makes it the ideal choice for dynamic sound manipulation but offers no control over meso and macro-formal aspects.

The first contribution of Mouja+ is a pipeline complementing these two methods with each other. By prompting textual instructions through carefully designed text scores, and by layering the generated files during playback in the guise of prompt jockeying, the composer gives shape to the piece. At the same time, through optional timbre transfer and latent manipulation with RAVE, one can intervene into the most minute timbral parameters as the piece unfolds. To easily combine Stable Audio and RAVE in live scenarios we developed a Graphical User Interface (GUI), through which one prompts textual instructions, plays back and mixes the generated cues, and optionally routes them to a series of RAVE models for real-time timbre transfer.

3.1.4 Graphical User Interface

Mouja+'s GUI (Fig.2) forwards to and download prompt from a remote server through a dedicated OSC application [3]. The added overhead of synthesizing the sounds on a remote server is justified by the need to reduce the time lag between prompting and the generation of the file, which is obviously a critical aspect in a live performance. By processing the prompt on a server with a high-performing GPU, we could sensibly reduce the time required for the generation of the sound files (around 20 seconds for a 47-second sample) compared with most consumer machines, even considering the time required to download the file into the client.

On the client GUI, one writes positive and negative text prompts, and defines length in seconds, number of steps, temperature, and buffer for the download. Once ready, the file generated by Stable Audio automatically uploads into the specified buffer on the client computer. The GUI allows one to view, play, process and mix up to eight audio files at the same time, that can be routed either directly to the Ambisonics encoder (and from there to the speakers), or (either entirely or partially) to RAVE for timbre transfer. When timbre transfer is applied, the performer can influence the resynthesis by manipulating the latents through a dedicated interface such as Stacco.

The GUI also allows one to actively control the position of up to four sound sources (one per buffer, for a total of four buffers) in the Ambisonics space, with the spherical coordinates of each source mapped to the position of one magnetic attractor in Stacco. When the sound sources are not actively controlled by the user, an algorithm handles their displacement along the circular perimeter of the dome, with different speeds and directions.

In experimenting with RAVE, we first trained a model with the same dataset (from the Freesound archive) used by Stable Audio Open. The results were not satisfactory, both in terms of sound quality and latent control, probably due to the high timbral variability and large size of the dataset, which seems to be at odds with RAVE's focus on small and timbrally consistent datasets. We thus selected a series of different models curated and trained by various members of the Intelligent Instruments Lab (IIL), including foley sounds, liquids, guitars, voices, percussions, magnets, organs, and birds. [4] This approach proved more effective, both in terms of extending the system's timbral possibilities, and of providing a fine control over the resynthesis through latent manipulation.

3.2 Performance

Mouja+ features ten text scores adapted from the Fluxus Performance Workbook [19], whose selection and adaptation I view as a practice-based investigation into how language use shifts when addressing AI systems. During the SWRL performance, scores were displayed on a large screen alongside real-time information indicating whether each prompt's synthesis was in process or complete, a deliberate choice that allowed me to play with the audience's cognitive dissonances between the textual instructions and the model's outputs.

The audio files were layered and structured into a tripartite form: two sections of three scores each, and one section of four scores. Score selection prioritized timbral and dynamic considerations, as well as textual relatedness, with the different text scores establishing semantic relationships to form a surreal and theatrical meta-narrative. The piece describes the entrance on stage of a piano, that gets interacted with in different ways (except playing it in a traditional way, as typical in Fluxus), then dismantled and sold along with the whole theatre.

The use of spatialisation techniques greatly influenced the selection of the Fluxus scores toward prompts that would evoke on stage the presence of large objects and acting bodies, whose sonic traces would vividly move around the audience. Audience members described the performance

[3] https://github.com/elgiano/gradio-osc

[4] https://huggingface.co/Intelligent-Instruments-Lab/rave-models

as evocative, due to the cognitive dissonance between the realism of the sounds, materialising the presences evoked by the text scores, and their absence on stage. Through the real-time displacement of the sources afforded by the pipeline, I could intensify this effect at will throughout the piece.

Section One opened with concrete sounds moving concentrically around the audience, and the noise of a heavy resonating body (a piano, as revealed by the score) being dragged into the stage. The text scores were three: 'Let piano movers carry the piano on stage,' 'Wash the piano, wax it and polish it well' (adapted from Nam June Paik), and 'Eat a juicy apple' (adapted from Bob Lens), the sequence suggesting the presence of one or more figures on stage, struggling to move a piano, washing it and finally enjoying a snack. As the files were generated and overlapped, I routed a part of the playback to RAVE, and controlled the resynthesis with Stacco, using a foley and water model to smooth out transitions and emphasize dynamics.

Section Two shifted from concrete sounds to pointillistic and evocative ones. The scores were 'Hammer nails into the piano keys' (adapted from Tomas Schmit), 'Dropping coins' (excerpt from George Maciunas Solo for Rich Man) and 'The sound of the stone ageing,' (adapted from Yoko Ono). As in the first section, I resynthesised the sounds with Stacco. Here, a crucial challenge was managing the model's tendency toward literalism (albeit an idiosyncratic one, as I discuss in section 4). While section one materialised objects and bodies on stage, this section's transition toward semantic ambiguity (e.g. the sound of a stone ageing) foregrounded the model's tendency to collapse uncertainty into overly simplistic interpretations. The section concluded with a solo passage using direct latent navigation on Stacco without resynthesis, momentarily stepping outside the text-to-audio paradigm to foreground the performer's embodied control over the latent space itself.

Section Three exploited the model's realism more deliberately, by conjuring on stage an open market, with unintelligible voices merging into a crowd that I dynamically displaced in the Ambisonics dome with Stacco. This reversed the circumscribed and closed space with that of an imaginary outside. Finally, the cacophony of voices, coins, pianos, deflated again into the crawling of a piano (the same piano that entered on stage at the opening) being heavily dragged out of the stage by imaginary actors. The text scores I used in this section were 'Performer sells the theatre' (adapted from Ben Vautier), 'Piano movers carry piano out of stage' (adapted from Nam June Paik), and, as a closure, Ono's 'The sound of the stone ageing.'

4. DISCUSSION

Through the first author's artistic practice on Mouja+, we gained multiple insights relevant to design practice with NAS, the use of NAS in combination with spatial audio, and best practices for designing text scores for text-to-audio generative models. In the discussion, we focus on the latter.

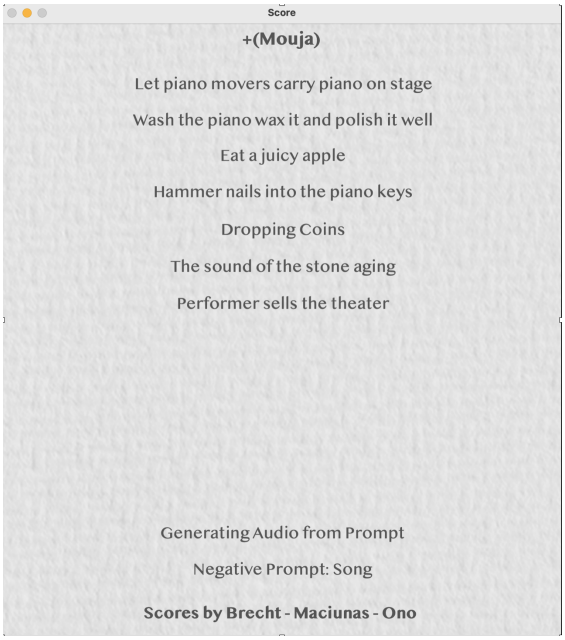

Figure 3. Mouja+ score combining adaptations of Fluxus text scores, as seen by the audience.

Figure 4. Performing Mouja+ at SWRL.

4.1 Language Games

In working on Mouja+, one aspect became inevitably apparent: when composing text scores for text-to-audio models, our use of language must change and adapt. We can examine this through Wittgenstein's notion of language games, which demonstrates how meaning in language is not fixed but emerges through contextual use [24].

Yoko Ono's Stone Piece offers a fitting example, as it plays with a radical ambiguity, by inviting the performer to 'take the sound of a stone ageing.' This is a contradictory statement on many levels. What is the sound of ageing? How can a stone age? And as a consequence, what does ageing mean? Like a haiku, this score introduces interpretative ambiguities and nuances that call for introspection. Yet, to an existential question requiring deep reflection and awareness of context of use, in other words of the language game being played, the model tends to default to a simplistic and literal interpretation: the sound of a stone being scratched.

From this reflection, it is clear how there exists a cer-

tain art to being literal with a text-to-audio model. This involves an adoption and use of natural language that accounts for the constraints determined by the model (perhaps better, by the encounter of a certain algorithm with a certain training data), that predates the performative contextualization. Tentatively discovering and adapting to the language game Stable Audio knows and plays best, and leveraging it in composing Mouja+ was perhaps the most challenging (but rewarding) process in working on this piece.

4.2 Compositional Strategies

To understand why the strategies described below were adopted, it is necessary to first clarify the artistic intent. To do this, we may draw from Umberto Eco's theory on openness and indeterminacy in contemporary artistic production. In *Opera Aperta*, Eco demonstrates how contemporary artists create open works by incorporating different forms of entropy, allowing for the contextual emergence of multiple meanings [14]. This implies walking a fine line between control and deliberate unpredictability, which is necessary to the emergence of poetic meaning.

Similarly to the examples discussed in 2.1, in Mouja+ Stable Audio's interpretative idiosyncrasies and statistical variability are viewed as yet another source of entropy, and the efforts in learning how to best design the prompts (the syntactic stratagems, but also the pipeline) represent yet another attempt to negotiate emergence and control.

In other words, the aim in Mouja+ was not one of turning a probabilistic system into a deterministic one, to know with a high degree of precision what sound will be generated given a certain prompt. Rather, it was to learn how to constrain the model's indeterminacy within a certain space of possibilities, to become familiar with how the system interprets language and anticipate, with reasonable confidence, how prompts might sound once rendered. The strategies presented in the next section were developed in order to achieve this balance. These may termed (i) *syntactic fine-tuning*, (ii) *mindful use of negative prompting*, and (iii) iterative *exploration of the training data*.

4.2.1 Syntactic Fine-tuning

I began working on Mouja+ by testing a large number of Fluxus scores. Once I found a score from which the model generated an artistically interesting output, I further experimented with it, first by working on the semantics of the score and then by changing the syntax. If semantic changes usually produced very different sonic worlds, by slightly changing the syntax (for instance by inverting the order two words or by adding or removing a preposition) I could finely stir the generation process towards a desired outcome. Metaphorically speaking, semantics changes felt like preset changes on a synth, and syntactic ones as fine-tuning a given preset. For instance, 'Let piano movers carry the piano out of the stage,' 'piano movers carry piano out of stage,' or 'eat juicy apple(s) during concert' and 'eat a juicy apple during concert' tend to be interpreted in similar ways by humans. However, by adding 'a' and removing '(s)' from the text score, I increased the consistency of the model, intended as the likelihood of generating similar results through repeated prompts using the same text.

4.2.2 Dataset Exploration

Stable Audio Open is trained on the Freesound archive, which is open and freely accessible online through a convenient search engine. This feature was highly relevant in shaping my prompts. I used Freesound's advanced search engine to explore text embeddings, typing keywords and selecting tags based on the scores I was experimenting with. One example is the word 'exit,' as per Brecht's score, which I experimented with extensively. As I realised by checking the keyword with Freesound's search engine, in the database 'exit' is often embedded in environmental recordings involving trains arriving at the station, opening and closing their doors, and leaving. This suggests that the statistical likelihood that prompts containing this word might incorporate sounds of moving trains is high, as was confirmed through successive tests with the model.

4.2.3 Negative Prompting

Once I understood how words were embedded in the original data, I typically decided whether to discard a score or use negative prompting to stir the system towards an expected outcome. In most cases, negative prompting was the most effective strategy. For instance, useful negative prompts for the word 'exit' were 'train' and 'bus,' 'station,' or 'soundscape,' which allowed avoiding sounds of public transportation coming and going. With negative prompts, I could effectively define niches of sounds by exclusion, counterbalancing some of the problems arising from the model's lack of context awareness. This was a critical aspect of Mouja+: in most of the scores I used negative prompts such as 'music,' 'piano music' or 'song' to avoid the generation of structured musical pieces of any genre, that would have hardly integrated with the rest of the performance.

4.3 Open Access

Of the three methods, dataset exploration deserves here further consideration. Checking the relationship between individual keywords (or clusters of words) and the dataset through Freesound's search engine was a particularly helpful strategy in adapting Fluxus scores to Stable Audio, one I systematically and iteratively applied on every single score. Crucially, the applicability of this method relies on the Freesound archive's openness and on the availability of an advanced search engine, two features that represent the exception rather than the norm in text-to-audio models and more broadly generative AI.

The academic literature pointing to the value of open access to datasets is rich. For instance, scholars advocate for FAIR (Findable, Accessible, Interoperable, Reusable) approaches to data collection and deployment as a way of mitigating bias [25], discuss the ethical relevance of data transparency [26], and criticise the concerning systemic lack of data documentation among AI companies [27]. In response to these pressing issues, researchers have implemented tools to trace lineages and audit text datasets [28]

and multimodal content, that allow querying a certain corpus, verifying whether and how data is licensed, or assessing diversity, inclusion, and biases.

The artistic research presented in this paper tackles this issue from a slightly different angle, showing how, instead of relying on tentative prompting approaches, we can more effectively refine our natural language instructions through the direct engagement with the training data. Of course, provided that the latter is accessible and can be effectively searched. This hints at how openness and accessibility are essential not solely in addressing critical issues around bias, data ownership and accountability, but also to facilitate and improve our engagement with generative AI within and beyond the arts.

5. CONCLUSION

In this paper, we investigated how text-to-audio models can serve as compositional and performative tools for experimenting with text scores. Our primary technical contribution demonstrates that by combining an asynchronous text-to-audio model (Stable Audio Open) with a real-time timbre transfer one (RAVE), performers gain control over macro-formal aspects and timbral navigation through NAS techniques. This hybrid approach extends text-to-audio generation beyond the limitations of prompt jockeying.

In Mouja+, we used Fluxus scores as probes to investigate our use of natural language when addressing AI systems rather than human performers. Drawing on Wittgenstein's language games, we found that effective communication with text-to-audio models requires adapting our strategies in designing the scores. Three practices proved particularly effective: syntactic fine-tuning of prompts, strategic use of negative prompting, and iterative exploration of the training dataset.

The last point in particular shows that dataset transparency and accessibility is crucial to modulate control and emergence, and thus to develop effective compositional workflows with generative systems. While current works emphasize the importance of openness and accessibility of datasets to address bias and ownership issues and ensure accountability, our work demonstrates that accessible training data also enables a more intentional and nuanced creative engagement with generative AI systems.

It is possible, and perhaps even likely, that the compositional and performative approach to AI co-creation via natural language we described will diffuse as NAS technologies increase in efficiency. But as our engagement with generative systems grows and evolves, affecting our lives on stage and outside of it, we should be mindful that, with AI, there is an art to being literal.

6. ETHICAL STANDARDS

All the RAVE models used in this work have been trained from open-source material or with the explicit consent of all parts involved, and are free to download and use. Stable Audio Open is trained on Creative Commons-licensed audio [15].

Acknowledgments

This research is supported by the European Research Council (ERC) as part of the Intelligent Instruments project (INTENT), under the European Union's Horizon 2020 research and innovation programme (Grant agreement No.101001848). We also thank SWRL space and Voldemārs Johansons for organising and hosting the performance.

7. REFERENCES

[1] V. Shepardson, J. Armitage, and T. Magnusson, "Notochord: a flexible probabilistic model for embodied midi performance," 2022. [Online]. Available: https://zenodo.org/record/7088404

[2] N. Privato, O. Rampado, and A. Novello, "A creative tool for the musician combining lstm and markov chains in max/msp," in *Artificial Intelligence in Music, Sound, Art and Design: 11th International Conference, EvoMUSART 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20–22, 2022, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 228–242. [Online]. Available: https://doi.org/10.1007/978-3-031-03789-4_15

[3] S. Yuditskaya, S. Sun, and M. Schedel, "Synthetic erudition assist lattice," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Shanghai, China, June 2021. [Online]. Available: https://nime.pubpub.org/pub/5oupvoun

[4] V. Shepardson and T. Magnusson, "The living looper: Rethinking the musical loop as a machine action-perception loop," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, M. Ortiz and A. Marquez-Borbon, Eds., Mexico City, Mexico, May 2023, pp. 224–231. [Online]. Available: http://nime.org/proceedings/2023/nime2023_32.pdf

[5] N. Privato, T. Magnusson, and E. T. Einarsson, "Magnetic interactions as a somatosensory interface," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, M. Ortiz and A. Marquez-Borbon, Eds., Mexico City, Mexico, May 2023, pp. 387–393. [Online]. Available: http://nime.org/proceedings/2023/nime2023_54.pdf

[6] J. Walshe, "The text score dataset 1.0," https://megansteinberg.com/wp-content/uploads/2022/01/a6702-thetextscoredataset1.0-jenniferwalshe.pdf, 2021, project documentation.

[7] S. L. Moan, "Machine listening: Machine learning and audio analysis," https://www.youtube.com/watch?v=_fpnAHoRSqU, Aug. 2020, youTube video.

[8] L. Russolo, *The Art of Noises*, 1986.

[9] J. D. Erickson, *Dada: Performance, Poetry, and Art*, ser. Twayne's World Authors Series. Boston: Twayne Publishers, 1984, vol. TWAS 632.

[10] K. Friedman, Ed., *The Fluxus Reader*. Chichester, West Sussex; New York: Academy Editions, 1998, free digital edition available from Swinburne University of Technology. [Online]. Available: http://hdl.handle.net/1959.3/42234

[11] P. Oliveros, *Anthology of Text Scores*. Kingston, NY: Deep Listening Publications, 2013, edited by Sam Golter and Lawton Hall.

[12] T. Funk, "A Musical Suite Composed by an Electronic Brain Reexamining the Illiac Suite and the Legacy of Lejaren A. Hiller Jr." 9 2018. [Online]. Available: https://indigo.uic.edu/articles/journal_contribution/A_Musical_Suite_Composed_by_an_Electronic_Brain_Reexamining_the_Illiac_Suite_and_the_Legacy_of_Lejaren_A_Hiller_Jr_/10781933

[13] J. Walshe and D. D. Roure, "Ai text scores," https://www.rncm.ac.uk/research/research-activity/research-centres-rncm/prism/prism-blog/ai-text-scores/, July 2021, pRiSM Blog, Royal Northern College of Music.

[14] U. Eco, *The Open Work*. Cambridge, MA: Harvard University Press, 1989, original work published as *Opera Aperta*, 1962.

[15] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," 2024. [Online]. Available: https://arxiv.org/abs/2407.14358

[16] Intelligent Instruments Lab, "Workshop: Exploring ai musicianship with dadabots," https://iil.is/news/dadabots-workshop, 6 2024, workshop held at Veröld – hús Vigdísar, University of Iceland, Reykjavik. [Online]. Available: https://iil.is/news/dadabots-workshop

[17] N. Privato, "Mouja: Experiencing ai through magnetic interactions," in *Proceedings of the Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction*, ser. TEI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3623509.3635328

[18] N. Privato, T. Magnusson, and E. T. Einarsson, "The magnetic score: Somatosensory inscriptions and relational design in the instrument-score," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'2023*, A. P. D. Ritis, V. Zappi, J. V. Buskirk, and J. Mallia, Eds. Boston, Massachusetts, USA: Northeastern University, 2023, pp. 36 – 44.

[19] K. Friedman, O. F. Smith, and L. Sawchyn, *The Fluxus Performance Workbook*. Swinburne, 2002. [Online]. Available: https://doi.org/10.25916/sut.26270992.v1

[20] A. Caillon and P. Esling, "Rave: A variational autoencoder for fast and high-quality neural audio synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2111.05011

[21] N. Privato, V. Shepardson, G. Lepri, and T. Magnusson, "Stacco: Exploring the embodied perception of latent representations in neural synthesis," pp. 424–431, September 2024. [Online]. Available: http://nime.org/proceedings/2024/nime2024_62.pdf

[22] N. Privato, G. Lepri, T. Magnusson, and E. T. Einarsson, "Sketching magnetic interactions for neural synthesis," in *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR'24*, P. Kocher, Ed. Zurich, Switzerland: Zurich University of the Arts, 2024, pp. 89–97.

[23] D. Stowell and M. D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," October 2013, arXiv preprint. [Online]. Available: https://arxiv.org/abs/1309.5275

[24] L. Wittgenstein, *Philosophical Investigations*, G. E. M. Anscombe, Ed. New York, NY, USA: Wiley-Blackwell, 1953.

[25] S. Raza, S. Ghuge, C. Ding, E. Dolatabadi, and D. Pandya, "Fair enough: How can we develop and assess a fair-compliant dataset for large language models' training?" 2024. [Online]. Available: https://arxiv.org/abs/2401.11033

[26] J. Hardinges, E. Simperl, and N. Shadbolt, "We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models," *Harvard Data Science Review*, no. Special Issue 5, may 31 2024, https://hdsr.mitpress.mit.edu/pub/xau9dza3.

[27] S. Longpre, R. Mahari, A. Chen *et al.*, "A large-scale audit of dataset licensing and attribution in ai," *Nature Machine Intelligence*, vol. 6, pp. 975–987, 2024. [Online]. Available: https://doi.org/10.1038/s42256-024-00878-8

[28] A. Piktus, C. Akiki, P. Villegas, H. Laurençon, G. Dupont, A. S. Luccioni, Y. Jernite, and A. Rogers, "The roots search tool: Data transparency for llms," 2023. [Online]. Available: https://arxiv.org/abs/2302.14035