Composing with an AI-Generated Sound Corpus: *Reflections on My Computer's Interpretation of Falling Down*

Stevie J. Sutanto
Faculty of Music
Universitas Pelita Harapan
Tangerang, Banten
stevie.sutanto@uph.edu

Abstract

This paper reflects on the creation of *My Computer's Interpretation of Falling Down*, a composition developed using an AI-generated sound corpus. Rather than using AI solely as a source of raw material, the work explores how generative models can also shape compositional structure and relationships between sounds. The process involved generating a corpus of motion-related sounds via text prompts submitted to a text-to-audio model, then organizing and sequencing those sounds through feature-based clustering. The result is a piece shaped through interaction—between language, system behavior, and listening. Motivated by curiosity about human–machine collaboration, the work explores how this approach might not only shape musical form but also reveal how a black-box generative model interprets a constrained topic through its underlying biases using language that is meaningful to humans, rather than adjusting abstract model parameters.

1 Introduction

My Computer's Interpretation of Falling Down explores the sonic qualities of objects in motion—falling, rolling, spinning, and sliding—through the lens of machine-generated sound. Rather than sourcing sounds from real-world recordings or synthesis, the material was developed using an AI-generated sound corpus constructed from a set of descriptive prompts. This approach examines how generative audio models, particularly text-to-audio models, which are typically employed to create single audio pieces, can be more fully integrated into the compositional process.

In this work, AI tools were tasked not only with producing sound material but also with contributing to the shaping of their organization and formal development. This ties into broader discussions in generative music research, which position AI systems not only as content generators but as potential co-creative agents (Singh et al., 2024). That sense of discovery—led partly by the system's own logic—was central to how this piece took shape.

2 Working with Text-to-audio Model

Text-to-audio models have quickly evolved, allowing artists to describe sound via language and receive audio renderings. However, their creative use is often limited to isolated tasks, like generating one-shot effects or filling sonic gaps. I aimed to push this further—viewing the AI-generated corpus not just as sounds but as a vital component in the compositional framework.

This curiosity relates to a broader question about the potential of generative models: how might their internal associations, unpredictable behavior, or imperfections be part of the compositional process? How can a text-to-audio system, when utilized creatively, contribute not only to a piece's timbral surface but also to its structural logic?

Proceedings of the 6th Conference on AI Music Creativity (AIMC 2025),
Brussels, Belgium, September 10th-12th

Recent research, such as that by Cherep et al. (2024), shows that AI-generated sound need not aim for realism; it can offer abstract renderings that evoke ideas rather than reproduce them. Similarly, Liu et al. (2025) highlight frameworks that position AI as a compositional collaborator, coordinating diverse audio elements in structured ways. This project fits within these explorations while focusing on a single, constrained world: objects in motion.

3 From Corpus to Composition

The generative process began with the use of a text-to-audio model accessed through ElevenLabs sound effects API. [1] In this project, I explored how such a commercial model could respond to a more focused set of physical-motion scenarios, aiming to build a unified and thematically constrained sound corpus.

To manage the diversity and specificity of the corpus, I created a simple python script that automatically generated descriptive prompts by combining variables such as material (e.g., wood, glass, metal), object size (small, medium, large), type of motion (falling, sliding, bouncing, rolling), and surface interaction (concrete, gravel, water, etc.). These phrases were then sent to the API, which returned a range of short audio clips based on the textual input. This automated approach enabled the model to respond with diverse sounds unified by a shared conceptual domain. Examples of the generated prompts include:

- "A heavy wooden object sliding across a concrete floor"
- "A small glass ball spinning to a stop on metal"
- "Something rubbery bouncing quickly on gravel"

The generated audio was analyzed using perceptual and acoustic features, including spectral centroid, brightness, duration, and envelope shape. Based on these features, I grouped the sounds into two broad categories: impact gestures (short, transient events) and motion gestures (sustained sounds like rolling or sliding).

To connect these categories, I used a KDTree structure to efficiently search for acoustic similarity. For each impact sound, the system retrieved nearby motion gestures in the feature space—those with comparable spectral and temporal characteristics. These pairings were not intended to be literal but were designed to create perceptual continuity: a sense that one sound might logically follow another, even if they originated from entirely different prompts.

The resulting sequences emerged through a balance of algorithmic association and listening-based curation. I treated the system's suggestions not as fixed solutions but as proposals—starting points for exploring connections, questioning assumptions, and shaping the piece through attention to what the material seemed to offer. In this way, the form of the composition was not imposed in advance, but gradually discovered through interaction with the corpus and the behavior of the system itself.

This method resonates with corpus-based (Schwarz, 2007) and AI-assisted composition practice, but with a key difference: the sound materials were not drawn from an archive of real-world recordings or instrumental samples. This process involved constructing the corpus from the ground up through interaction with a generative model, serving as both a dataset and a creative environment.

4 Challenges

Working with AI-generated sound corpora presented a number of challenges—some technical, others more conceptual. One recurring difficulty was the unpredictability of the model's output. While the structure of the prompts provided a degree of control, the results were not always consistent or clearly aligned with the intended motion. Some sounds arrived overly abstract, while others felt distant from the physical behaviors I had in mind. At first, this unpredictability felt like a limitation, but over time I began to view it as part of the material's expressive range. Many of the textures that became essential to the final piece emerged precisely from these unexpected responses.

There was also a question of authorship. While I designed and curated the system, many of the specific sonic decisions—the choice of gestures, their order, their timing—were shaped by the model's

[1] https://elevenlabs.io/sound-effects

interpretations and clustering. Rather than seeing this as a loss of control, I approached it as a kind of co-composition, where the system's proposals acted as a stimulus for creative response.

5 Reflection

This composition is a small step in exploring how AI-generated sound corpora might be used not only as a source of material but also as a method for shaping musical form. By inviting a generative system to take part in both sound production and structural organization, I hoped to test what kinds of musical thinking might emerge.

The result is a piece shaped as much by listening as by designing—listening to the outputs of the model, to the relationships between sounds, and to the way form began to take shape through these interactions. Rather than aiming for a demonstration of technical novelty, the work leans into the uncertainties of the process: the occasional mismatches, the unlikely pairings, the surprising continuities that emerged through the system's internal logic.

While still in an exploratory stage, I see this approach as a step toward more integrated uses of AI in sound composition—where generative models contribute not only to what we hear but also to how we imagine and shape the spaces between sounds. At the same time, this method offers a way to explore how a black-box generative model interprets specific topics and reveals underlying biases—through language that is meaningful to humans, rather than by tuning abstract model parameters.

References

Cherep, M., Singh, N., and Shand, J. (2024). Creative text-to-audio generation via synthesizer programming. *arXiv preprint arXiv:2406.00294*.

Liu, X., Zhu, Z., Liu, H., Yuan, Y., Huang, Q., Cui, M., Liang, J., Cao, Y., Kong, Q., Plumbley, M. D., et al. (2025). Wavjourney: Compositional audio creation with large language models. *IEEE Transactions on Audio, Speech and Language Processing*.

Schwarz, D. (2007). Corpus-based concatenative synthesis. *IEEE signal processing magazine*, 24(2):92–104.

Singh, N., Mishra, M., and Machover, T. (2024). AI for Musical Discovery. *An MIT Exploration of Generative AI*. Publisher: MIT.