
# Neural audio instruments: epistemological and phenomenological perspectives on musical embodiment of deep learning



Citation for the original published paper (version of record):
Zappi, V., Tatar, K. (2025). Neural audio instruments: epistemological and phenomenological perspectives on musical
embodiment of deep learning. Frontiers in Computer Science, 7.
http://dx.doi.org/10.3389/fcomp.2025.1575168

frontiers | Frontiers in Computer Science

TYPE Hypothesis and Theory
PUBLISHED 20 August 2025
DOI 10.3389/fcomp.2025.1575168



OPEN ACCESS

EDITED BY
Stefania Serafin,
Aalborg University Copenhagen, Denmark

REVIEWED BY
Alexander Refsum Jensenius,
University of Oslo, Norway
Hamza Bayd,
EuroMov Digital Health in Motion IMT Mines Ales, France

*CORRESPONDENCE
Victor Zappi
✉ v.zappi@northeastern.edu

RECEIVED 11 February 2025
ACCEPTED 23 July 2025
PUBLISHED 20 August 2025

CITATION
Zappi V and Tatar K (2025) Neural audio instruments: epistemological and phenomenological perspectives on musical embodiment of deep learning.
*Front. Comput. Sci.* 7:1575168.
doi: 10.3389/fcomp.2025.1575168



# Neural audio instruments: epistemological and phenomenological perspectives on musical embodiment of deep learning

Victor Zappi[1]* and Kıvanç Tatar[2]

[1]Department of Music, College of Arts Media and Design, Northeastern University, Boston, MA, United States, [2]Data Science and AI Division, Computer Science and Engineering Department, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden

Neural Audio is a category of deep learning pipelines which output audio signals directly, in real-time scenarios of action-sound interactions. In this work, we examine how neural audio-based artificial intelligence, when embedded in digital musical instruments (DMIs), shapes embodied musical interaction. While DMIs have long struggled to match the physical immediacy of acoustic instruments, neural audio methods can magnify this challenge, requiring data collection, model training and deep theoretical knowledge that appear to push musicians toward symbolic or conceptual modes of engagement. Paradoxically, these same methods can also foster more embodied practices, by introducing opaque yet expressive behaviors that free performers from rigid technical models and encourage discovery through tactile, real-time experimentation. Drawing on established perspectives in DMI embodiment literature, as well as emerging neural-audio-focused efforts within the community, we highlight two seemingly conflicting aspects of these instruments: on one side, they inherit many "disembodying" traits known from DMIs; on the other, they open pathways reminiscent of acoustic phenomenology and soma, potentially restoring the close physical interplay often missed in digital performance.

KEYWORDS

**neural audio, digital musical instruments, neural audio instruments, embodied interaction, music performance, artificial intelligence, deep learning, latent space**

## 1 Introduction

Physical, sensory and cognitive engagement are fundamental aspects of musical instrument design, especially during the exploration of novel sound technologies. In the literature, the term *embodiment* is often used to refer to these intertwined aspects of the experience of music making. For example, the embodiment of a guitarist can be expressed as the holistic experience of using hand movements (physical, bodily motion) to press strings, feeling the instrument's body and its vibrations (sensory experience), and simultaneously interpreting chords and melodies (cognitive processes). And, clearly, the design of the instrument itself has a key role in fostering such a nuanced experience. The very concept of embodiment, though, goes beyond the domain of music and musical instruments. Its roots can be found in cognitive psychology (Varela et al., 2017) and philosophy (Merleau-Ponty et al., 2013), and its influence is at the base of theories that

defined some of the most important paradigms in modern design and human-computer interaction (e.g., Dourish, 2001; Kirsh, 2013; Höök, 2018).

In the context of musical instrument design, there exist several key factors that contribute to embodied musical interaction. These include: affordances embedded in the design, including physical structure (Keebler et al., 2014), dimensions (Mice and McPherson, 2021), as well as hidden modalities of interaction (Zappi and McPherson, 2018); extent (Caramiaux et al., 2015) and familiarity (Essl and O'Modhrain, 2006) of the sensorimotor experience leveraged by the instrument; emotional (Leman and Maes, 2014) and cultural connection (Tragtenberg et al., 2024) with the instrument and the music that it makes possible. The primary engine for these and many other factors tends to be the chosen design methodology and the specific rationales it puts in the spotlight—e.g., *soma design* and the focus on sensation and movement (Höök, 2018), human-computer interaction *entanglement* and the mutual influence between technology,[1] behaviors, culture and social norms (Morrison and McPherson, 2024). However, at a lower level it is often the technology that enables, enhances or in some cases even hinders the targeted embodied experiences.

In the case of acoustic-instrument design, "technology" refers to craftspeople's mastery of physical materials and their acoustic affordances. As discussed by Magnusson (2009), "various solutions are therefore introduced [by acoustic-instrument designers] to adapt to or extend the scope of the human body with the use of levers, keys and other mechanisms," grounding design in bodily experience *through* technology. When shifting to the domain of digital musical instruments (DMIs), a different set of technologies become available. Yet, many of the processes that characterize traditional craftsmanship remain present and get integrated with elements from physical computing and interaction design (Jordà, 2005). In this context, the first technologies to become fundamental vehicles of embodiment are sensors (Küssner et al., 2014), multimodal feedback apparatuses (Höök et al., 2021; Leonard et al., 2014) and highly responsive audio interaction platforms (McPherson, 2017), all harnessed differently by various theoretical frameworks to realize and assess embodied musical interaction.

As new technologies emerge, the horizons of instrument design expand, extending the possibilities for embodied musical interaction. Without a doubt, recent advancements in artificial intelligence (AI) are playing an increasingly significant role in how we craft DMIs and how we conceptualize our epistemological and phenomenological relationships with them. The most evident case involves musical interaction with intelligent musical agents (Tatar, 2019; Erdem, 2022). However, there is a growing interest in reimagining the embodiment of instruments that incorporate AI and machine learning algorithms for control and synthesis purposes, yet are not autonomous (e.g., Fiebrink and Sonami, 2020; Pelinski et al., 2022).

This work belongs to the latter research stream.[2] Over the last five years, we have focused our efforts on integrating deep learning technologies into DMI design for sound synthesis and processing—we refer to such devices as *neural audio* DMIs. Initially, our exploration was technically driven, centered on building new neural-audio models and deploying them on real-time platforms. Then, through both successes and failures, this hands-on work gradually expanded into a broader inquiry. Our deep dive into the domain of neural audio DMIs has nurtured new perspectives on how to use this evolving technology, with a specific focus on *embodiment* in relation to the better-studied domain of "canonical" DMIs (i.e., those that utilize non-neural audio technologies and have been explored in terms of their embodied interactions). In this work, we aim to share these perspectives, which we believe will be valuable for designing future DMIs that incorporate neural audio synthesis and processing, as well as for advancing the general development of this technology.

We firmly believe that now is the right time to initiate this discussion. Neural audio DMIs are beginning to emerge from the more abstract research challenges of machine learning, statistics and mathematics; these fields collectively provide much of the knowledge base and tools needed to develop neural models that understand, generate and manipulate audio, fostering the exploration of innovative musical paradigms. As of today, only a handful of fully fledged neural audio DMIs have been presented in the literature (e.g., Privato et al., 2023; Diaz et al., 2023) and companies and creators are just starting to invest in neural audio plug-ins and applications that can be used for musical performance (see, for example, Neutone,[3] Combobulator,[4] or Guitar Rig[5]), rather than solely for generating new sounds and novel production workflows. Yet, the number of research environments and tools (Caillon and Esling, 2021; Tatar et al., 2023), bespoke systems and experiments (Nercessian et al., 2023; Shepardson et al., 2024) that explore and extend the potential of neural audio is quite impressive (Hayes et al., 2024), especially considering how young this technology is (Hoopes et al., 2024). This extensive body of work tends to focus on output quality, computational efficiency and other technical metrics that lay the foundation for sonic and musical interaction in the digital domain. However, to define a musical instrument and its embodied playing experience, it is necessary to move beyond technical specifications and functional goals, and aim for a deep understanding of the epistemological and phenomenological aspects of the musician-instrument relationship (McMillan and Morreale, 2023), especially when this relationship is mediated by groundbreaking technology.

The rest of this article is organized as follows. The next section delves deeper into the context of DMI embodiment,

1 Throughout this article we use the term "technology" in its broad, classical sense, combining *techne* (craft knowledge, tacit know-how) with the material artifacts that manifest that knowledge.

2 Along these lines, throughout this article we reserve the word "instrument" for a physical or virtual device whose sound is primarily enacted by a human performer. Systems deliberately designed for a high degree of autonomy (often described as *interactive music systems* or *agents*) are mentioned only for context and lie outside the present focus.

3 https://neutone.ai/fx

4 https://datamindaudio.ai/

5 https://www.native-instruments.com/en/products/komplete/guitar/guitar-rig-7-pro/

surveying canonical instruments and laying out pertinent theories of embodied musical interaction. Section 3 turns to musical AI, tracing how algorithmic and autonomous approaches connect to mid-twentieth-century conceptions of "organized sound." We then define neural audio instruments in Section 4, explaining their key architectures and real-time potentials. Building on that, Section 5 examines how these instruments can transform yet also inherit embodiment challenges known from canonical DMIs. Finally, Section 6 consolidates our insights and offers practical recommendations for designers seeking to integrate neural audio into DMIs in ways that support strong embodiment and creative performance.

## 2 DMI embodiment

Embodiment can be *strong* when playing DMIs. We believe it is important to emphasize this perspective at the outset to better contextualize the subsequent sections of this article within the literature on DMIs and neural audio. Numerous studies have discussed how the gestural and perceptual experiences involved in playing DMIs can foster a physical and emotional connection with the instrument, achieved through diverse combinations of underlying technologies and mapping strategies (Malloch and Wanderley, 2017), as well as design metaphors (Fels, 2004). To fully grasp the extent of embodiment described in these works, we argue that it is useful to analyze them through the lens of *flow*.

The concept of flow, introduced by Csikszentmihalyi (1975) as "the holistic sensation that people feel when they act with total involvement," provides a valuable framework for understanding the embodied experience of playing DMIs. In other words, flow is a mental state characterized by complete absorption and effortless engagement, leading to optimal performance. Since its introduction, flow has been studied in various contexts, with music-making being among the most prominent (Wrigley and Emmerson, 2013; Chirico et al., 2015; Habe et al., 2019). Crucially, the phenomenon of flow is frequently experienced when playing DMIs, as these interactions inherently engage both physical and cognitive dimensions (Fels, 2004; Zamborlin, 2015; O'Modhrain, 2011; Nick, 2007). Nash's doctoral research Nash (2011) explored how composers and performers attain flow with music production software by analyzing more than 1,000 creative sessions. Post-task surveys, adapted from the *Dispositional Flow Scale-2* (Jackson and Eklund, 2002), revealed recurrent episodes of deep focus, loss of self-consciousness and a distorted sense of time. Complementing these quantitative findings, Reed et al. (2022) focused on performer experiences with more idiosyncratic DMIs. Through micro-phenomenological interviews, the authors gathered data showing how, under common live performance conditions, performers of augmented or custom-built instruments entered similarly effortless states. Anecdotal evidence from practitioners echoes these academic results; the electronic artist Tim Exile, for instance, openly discussed the importance of flow during his creative practice[6] and even built a custom instrument—called the *Flow Machine*—explicitly to sustain that state throughout performance.

Although flow and embodiment are distinct concepts, they are closely interconnected in the context of playing DMIs. Statements like that of Fels (2004), who observed that "an embodied interface allows expression to flow," suggest a strong relationship between the two. Expanding on this connection, Armstrong (2006) described the experience of flow as "a heightened sense of embodiment." Numerous accounts in the DMI literature reinforce the idea that musical flow is deeply dependent on the embodiment of the instrument. Therefore, the frequent attainment of flow in DMI performance affirms that strong embodiment is inherent to playing these instruments.

Building on this intrinsic link between flow and strong embodiment in DMI performance, we can explore how embodiment is central to experiencing the fundamental phenomenological aspects of flow: effortlessness, absorption and skill development through challenge. For instance, Fels (2004) argues that to enable expressive musical performance without cognitive strain, the DMI—or "interface," as Fels calls it—must become an extension of the player's body. From an enactive perspective, as embraced by Armstrong and later analyzed by Magnusson (2009), this embodiment can reach additional phenomenological modes. The DMI can become a medium for a hermeneutic relationship with the world, facilitating deeper understanding and interaction. Embodiment can even extend into epistemological modes. When the DMI acts as "a conveyor of knowledge used by an extended mind [...] that affords cognitive offloading," as described by Magnusson (2009), it enables what he refers to as *virtual embodiment*. This concept highlights how instruments can mediate and augment cognitive processes, further intertwining embodiment with the flow experience.

Discussing the player's absorption into musical action during flow, Armstrong (2006) describes a feeling that is "directly lived and experienced." This resonates with more recent DMI research that studies embodiment through the lens of Shusterman's somaesthetics (Shusterman, 2008). Specifically, both Tapparo and Zappi (2022) and Martinez Avila et al. (2020) present examples of DMI designs that enhance the player's *bodily awareness* of the physical processes involved in music making. In these designs, the actions that drive sound production/control are deautomatized, allowing the player to become fully immersed in their interaction with the DMI. As Martinez Avila et al. (2020) remark, this immersion facilitates the instrument becoming an extension of the body, echoing Fels' observations and underscoring how the fundamentals of flow and embodiment are intertwined.

Achieving a state of flow also requires the musician to possess above-average skills and to confront meaningful challenges (Csikszentmihalyi, 1975; Wrigley and Emmerson, 2013). The development of skills is a key part of the process of DMI embodiment[7]—encompassing both intellectual and perceptual-motor abilities (Gurevich and Fyans, 2011). Oore (2005) and Cannon and Favilla (2012) describe skill development as a journey involving a gradual shift from focusing on technical aspects of

6 https://timexile.medium.com/endlesss-is-beginning-13628c5030e7

7 In the case of acoustic instruments, the development of highly refined skills is also the path to virtuosity. Yet, interestingly, Morreale et al. (2018) found that with DMIs virtuosity is not always the musician's primary aim.

the interface to cultivating a deeper understanding of the DMI's expressive potential. Thus, embodiment is not immediate but unfolds progressively as the performer gains proficiency, reflecting the necessity of mastering skills before reaching a state of musical flow. Furthermore, Fels (2004) adds that "as a player continues to practice, the requirement for embodiment [...] need[s] to change to keep the player's interest." This ongoing need for viable challenges—which Wessel and Wright (2002) argue is sustainable only if the instrument lacks a performance *ceiling*—has been emphasized by numerous musicians and scholars studying DMI design and embodied interaction (e.g., Jordà, 2004; McDermott et al., 2013; McMillan and Morreale, 2023). Collectively, these accounts highlight how flow and its prerequisites depend on embodiment, and how the strength of embodiment makes flow possible when playing DMIs.

A significant strand of DMI research centers on *gestures* (Tanaka, 2010; Jensenius, 2014), here meaning the performer's bodily actions that are algorithmically mapped to sound synthesis (Hunt and Wanderley, 2002). Rich gestural interaction is often credited with fostering flow (Moral-Bofill et al., 2022) and a strong sense of embodiment (Jensenius and Wanderley, 2010). Yet, in the context of DMIs, such experiences are not confined to overt movement and limited gestures do not necessarily equate to a lack of flow and of embodiment. Indeed, a wide variety of DMIs exist that do not heavily rely on gestures and are described by Magnusson (2009) as being "essentially of a symbolic nature." Nevertheless, these instruments are frequently associated with states of flow during performance. As Green (2011) notes, symbolic mediations, while not unique to digital systems, do not inherently disrupt embodied flow. An indicative example is live coding, a digitally mediated musical practice extensively studied for the specific cognitive load that musicians manage during performance (Nick, 2007; Sayer, 2016). In live coding, gestures and movements are less than special; they appear mundane, stemming from everyday computer interactions rather than traditional musical expressions (Gurevich and Fyans, 2011). Nonetheless, we argue that within the context of DMI performances, there is perhaps no clearer example of being *in the flow* than an expert live coder during a show. Anyone with even cursory programming experience can immediately sense the laser-focus required to juggle algorithms, syntax and groove in front of a cheering crowd. One stray bracket and the whole party crashes!

This observation is supported by numerous studies in the literature that have systematically examined both the skills and music-making experiences of expert live coders (e.g., Brown and Sorensen, 2009; Roberts and Wakefield, 2018; Sayer, 2016), as well as research focused specifically on audience perceptions of performers' cognitive states. For instance, Burland and McLean (2016) investigated live coding performances in contemporary contexts, such as algoraves and stage setups featuring large-scale projections of the performer's screen. Their findings highlight frequent audience perceptions that performers enter what appears to be a flow state, with the inherent unpredictability and risks of real-time coding amplifying the audience's impression of coders being deeply "in the zone." Similarly, Roberts and Wakefield (2018) describe instances where performers become so deeply immersed in the creative act that they seem to transcend self-consciousness, fully absorbed in the "authentic moments of decision" that unfold live on stage.

This points to a profound yet distinct type of embodiment, one still grounded in gestures but of a different nature, connecting physical actions with cognitive and intellectual engagement. Building on Sayer (2016), live coders translate fleeting musical ideas into executable code through rapid keystrokes, so minimal physical actions instantly reshape both the program and the resulting sound. Such technology-mediated actions fit squarely within the expanded gesture taxonomy proposed by Jensenius and Wanderley (2010), which explicitly accommodates interactions whose primary locus of expression is not direct sound production, but the real-time control and reconfiguration of symbolic materials. Viewed through this lens, embodiment is no longer restricted to conspicuous instrumental motions; it also encompasses the performer's engagement with abstract representations and their immediate sonic consequences. This perspective extends beyond live coding to a wide range of DMIs in which embedded algorithms are coupled with idiomatic physical interaction. User studies on novel instruments—for instance, those by Gurevich and Fyans (2011) and Zappi and McPherson (2018)—show that metrics such as playing technique, appropriation and audience perception can still reveal rich layers of embodied experience, even when the performer's actions are partly symbolic or computational in nature.

## 2.1 The limits of digital embodiment

Despite the strong sense of embodiment that *can* arise when playing DMIs, performers and scholars still often describe these instruments as *less embodied* than traditional acoustic and electric instruments. Here "less embodied" refers to an experiential imbalance. As outlined earlier in this section, many DMI designs foreground symbolic or cognitive work while offering comparatively little continuous bodily/sensory engagement. Because embodiment, in our definition (see Section 1), stems from the intertwining of physical, sensory and cognitive engagement, any design that privileges one pole at the expense of the others risks being perceived as less *holistic*. The next paragraphs analyze specific design contingencies—including control–synthesis split, resolution bottlenecks, haptic absence and symbolic mappings—that the literature links to this skewed experience.

Hunt and Wanderley (2002) highlight that "digital musical instruments are unique in that the control surface is often completely separate from the sound synthesis engine, enabling arbitrary mappings between performer actions and sound synthesis parameters." While this separation offers unprecedented possibilities, it can diminish the immediacy and physicality associated with traditional instruments, where action and sound are inherently linked through the instrument's material structure. In other words, DMI design can lead to more abstract, less embodied interactions, ultimately detaching the performer from the physical instrument (Bang and Fdili Alaoui, 2023; Magnusson, 2009). As Gurevich (2014) observes, "the separation of human action from the sound-producing mechanism limits the potential

for skilled sensorimotor engagement," reinforcing the notion that DMI design innately impacts the performer's embodied relationship with the instrument. Furthermore, Gurevich adds that "the blame [for the loss of sensorimotor skill] is assigned to the very nature of digital systems but also to their designers," for an *effortless* interaction is often regarded—or used to be regarded—as a "cardinal virtue" of computers (Ryan, 1991; McDermott et al., 2013).

The absence of haptic feedback in many DMIs further exacerbates this issue. Research indicates that the lack of tactile sensations–such as vibrations or pressure–can significantly impact the user's experience and engagement with the instrument (O'Modhrain, 2001; Young et al., 2018). For instance, when users interact with intangible DMIs that require mid-air gestures, they often report a feeling of disconnection from the sound being produced; this is due to the sound not emanating from a physical object that can be touched or manipulated (de Lima Costa et al., 2020) and to the action–sound timing often feeling off (Dahl, 2014). This contrasts sharply with acoustic instruments, where the physicality of the instrument provides immediate sensory feedback, reinforcing the embodied experience of music-making. And, beyond immediacy, subtlety plays a crucial pedagogical and expressive role in the acoustic domain. Musicians learn to shape sound through fine motor adjustments guided by tactile and kinesthetic feedback, developing a heightened awareness of the body–instrument coupling over time (O'Modhrain and Gillespie, 2018; Saitis et al., 2018). These phenomenological nuances are often muted when DMIs are built around specific technologies, including: virtual/mixed reality (Mäki-Patola et al., 2005; Serafin et al., 2016), three-dimensional user interfaces (Berthaut et al., 2011) and non-invasive input modalities (Knapp and Lusted, 1990; Ivanyi et al., 2023). In such cases, the absence of direct tactile references and the reliance on mediated interactions can heighten the sense of abstraction and constrain the modalities of embodiment.

The relatively limited embodiment in DMIs can also be attributed to the constraints imposed by their control technologies. As Moore (1988) highlights, the finite resolution and timing inaccuracies of protocols like MIDI introduce a *bottleneck* in communication between the performer and the instrument, effectively narrowing the range of expressive possibilities. Although one might assume that advancements in control, sensor and transmission technologies over the past three decades have resolved these issues, recent literature continues to highlight open challenges in both control (Zappi and McPherson, 2014; Jack et al., 2017b) and temporal resolution (McPherson et al., 2016). These persistent limitations are not merely technical but fundamentally affect the phenomenological experience of interacting with DMIs, as performers must often adjust their actions to accommodate the system's restricted input bandwidth. Unlike traditional instruments, which rely on a continuous, finely timed *action–sound coupling*—that is, a direct link between the performer's gesture (control action) and the resulting sound—DMIs must often thread their interaction through these digital bottlenecks (Gurevich, 2014), leading to a more mediated and less intuitive experience.

The consequence of these contingencies is that users may find themselves engaging with DMIs in a more cognitive and less instinctive manner. This shift increases reliance on learned behaviors and mental models rather than on intuition and tacit knowledge (Magnusson, 2009). Even through the lens of disciplines and philosophies such as embodied cognition and somaesthetics—which tend to emphasize embodiment as a pervasive aspect of human experience—it is acknowledged that a reduction in instinctive engagement and tacit knowledge leads to a corresponding decrease in the depth of embodiment. An interesting and conspicuous marker of this diminished sense of embodiment in comparison to traditional instruments is posture. The expression of musical embodiment during performance is significantly influenced by posture, which acts as a conduit for channeling musical ideas through the body and the instrument (Godøy, 2017). Paradigms like movement-based wearable instruments or body-centered design approaches intentionally support more natural posture and individualized ergonomics when using digital technologies (Cavdir and Dahl, 2022; Kirby, 2023), and several DMI musicians are committed to researching the "ideal" setup (for example, Tim Exile and his Flow Machine, mentioned earlier). Nonetheless, it is not uncommon to see DMI performances characterized by awkward and uncomfortable setups, with musicians bending over tables that are too low, too small or otherwise ill-suited to accommodate the components of the instrument and facilitate seamless interaction. Compounding this issue is the lack of substantial research aimed at defining appropriate posture for the unique and idiosyncratic nature of DMIs. This gap exists both at the stage of instrument design (Martinez Avila et al., 2023; Mice and McPherson, 2021) and within the realm of performance practice (Oore, 2005), further limiting the potential for an embodied connection between the performer and DMIs.

## 2.2 Fostering DMI embodiment

In the literature, researchers and designers have suggested numerous ways to address the challenges posed by the limited embodiment of DMIs and to improve performers' embodied connections with these instruments. Their contributions provide critical insights into how we can bridge the gap between performers and digital instruments while maintaining the flexibility and innovation that DMIs afford.

Integrating haptic feedback remains crucial for addressing the detachment performers often experience. For instance, Jack et al. (2017a) demonstrates how vibrotactile feedback can sustain the performer's connection to the instrument by directly linking gestural input with dynamic sonic responses. Building on this, self-sensing haptic actuators (Davison et al., 2024) offer real-time feedback through collocated sensing and actuation, further enhancing the musician's sensory connection by coupling physical interaction with nuanced control. Additionally, focusing on materiality and physicality–such as through interactive surfaces (Bang and Fdili Alaoui, 2023) or even passive haptics (Çamcı and Granzow, 2019; Lindeman et al., 1999)–can counteract the abstraction introduced by mid-air or virtual interaction paradigms, ensuring that instruments feel like natural extensions of the body.

As highlighted by Wessel and Wright (2002) over two decades ago, overcoming the control bottleneck requires designers to move beyond traditional protocols like MIDI and adopt systems capable of higher resolution and tighter temporal synchronization. Isochronous control-audio streams (Neupert and Wegener, 2019) offer a compelling solution by ensuring real-time synchronization between control data and audio streams, enabling more precise and responsive interactions. Real-time processing platforms like Bela (McPherson, 2017) build on this foundation, combining ultra-low latency with scalable, nuanced sensor input (Jack et al., 2017b) in an embedded form factor. By leveraging these advancements, DMIs can sustain the intricate relationship between gesture and sound that is central to traditional instruments, fostering deeper musical embodiment and expressive potential.

Shifting the paradigm from effortless interaction to a more deliberate, skill-based design ethos can further enhance embodiment. As Gurevich (2014) argues, fostering skilled sensorimotor engagement requires intentional design choices that embrace complexity and resist oversimplification. Tools and practices from Dalcroze eurhythmics (Bang and Fdili Alaoui, 2023) and somaesthetics (Hayes, 2022; Tapparo and Zappi, 2022) can inspire designs that prioritize embodied learning, ensuring that DMIs cultivate, rather than diminish, the performer's sensorimotor skill and tacit knowledge.

# 3 A new frontier: musical AI

Neural audio instruments represent just one manifestation of a broader and rapidly evolving area known as *musical AI*. While this article eventually narrows in on neural audio instruments and their unique implications for embodied musical interaction, it is useful first to establish how *musical AI* emerged, evolved and connected to earlier traditions of generative music and algorithmic composition. By clarifying its historical lineage and surveying existing literature, we can better situate the novelty of neural audio within this diverse landscape of computational musical systems.

## 3.1 From emergence to autonomy

Musical AI has gained significant visibility over the past two decades (Ma et al., 2024; Carnovalini and Rodà, 2020; Briot and Pachet, 2020; Tatar, 2019; Herremans et al., 2017), encompassing diverse computational approaches for creating and manipulating musical materials in *autonomous* or semi-autonomous contexts. These approaches can draw on "good-old-fashioned AI" (Russell and Norvig, 2010) (logic- and rule-based systems), statistical or Markov models (Begleiter et al., 2004), evolutionary algorithms (Sivanandam and Deepa, 2007), multi-agent systems (Wooldridge, 2009), or deep generative methods (Tomczak, 2022; Goodfellow et al., 2016)—all supporting a variety of musical practices, from symbolic composition to sound installations. Although these methods differ in technological detail, they share a unifying vision of using computational machinery to undertake tasks traditionally reserved for human composers and performers.

Despite the recent proliferation of machine learning in music, the idea of autonomous musical processes predates modern computation by centuries (Fowler, 1967). Generative art, as Galanter (2003) explains, involves an artist establishing rule sets or procedures (ranging from natural language instructions to mechanical devices) that operate with "some degree of autonomy," ultimately producing an artwork. Generative music builds on this principle, applying algorithmic methods to assemble musical material. In *Algorithmic Composition*, Nierhaus (2009) frames "composing by means of formalizable methods," suggesting that a machine is not strictly required for such processes if composers themselves employ if-and-then rules or other systematic procedures. Over time, researchers have sought increasingly sophisticated computational systems to automate or augment these generative processes, crystallizing the subfield of musical AI.

Various reviews showcase the expanding methodologies and applications in musical AI. Fernández Rodríguez and Vico Vela (2013) offered a focused perspective by linking AI in music to computational creativity, whereas Herremans et al. (2017) proposed a functional conceptual framework that grounds musical AI in more conventional music-theory. Tatar and Pasquier (2019) surveyed 78 musical AI systems through a multi-agent lens (Wooldridge, 2009), while Carnovalini and Rodà (2020) conducted a meta-review of articles covering "algorithmic composition," "computational creativity" and related keywords. Meanwhile, Briot and Pachet (2020) zeroed in on deep learning-driven systems and Ma et al. (2024) examined the rise of large, multipurpose or "foundation" models for musical tasks. Despite differing emphases, ranging from symbolic composition to interactive sound installations, these surveys underscore the interdisciplinary and evolving character of the field, as well as its potential to challenge preexisting boundaries in music technology.

## 3.2 20th century music, organized sound and beyond

If musical AI explores how computational systems autonomously generate or transform musical materials, the question of *what* those materials can be was radically reimagined in the mid-twentieth century. Composers like Luigi Russolo, Edgard Varèse and Karlheinz Stockhausen set the stage for a conceptual leap from tightly constrained musical resources to the notion that *any sound* is musical material.

Luigi Russolo's *The Art of Noises* (Russolo, 1913) stands out as an early manifesto contending that the industrial noises of modern cities constitute legitimate musical material. In a rapidly urbanizing world, Russolo saw the roar of motors, machinery and the "eddying of water and gas in metal pipes" not as mere background clamor, but as potential compositional elements. His audacious call to "substitute for the limited variety of timbres [in the orchestra] the infinite variety of timbres in noises" signaled a major expansion of the sonic palette, foreshadowing later electronic—and ultimately computational–approaches.

Building on this expanded view of sonic material, Edgard Varèse famously defined music as "organized sound" (Varèse and Wen-chung, 1966). His essay, *Liberation of Sound*, repositions the composer as a "worker in rhythms, frequencies, and intensities," challenging older assumptions hinged on emotional

affect, harmony, or melody. By advocating all sound as potential raw material, Varèse dismantled conventional boundaries of musical "beauty," embracing new forms of expression, including early electronic music. This departure from restrictive definitions resonates today in AI-driven systems that treat a vast array of recorded or synthesized sounds as input for algorithmic manipulation.

Following the notion of organized sound, the materiality of sound has become a prominent thread in twentieth century sound studies. Pierre Schaeffer's notion of "reduced listening" (Schaeffer, 1964) proposed focusing on sound "as is," without reference to its other aspects—such as semantics—highlighting sonic materiality and its potential musical qualities. R. Murray Schafer further proposed a taxonomy of sound types (Schafer, 1977), which propelled a research thread on categorizations of sound. Both Schaeffer's and Schafer's proposals were critically analyzed later in the twentieth century (Demers, 2010), enriching the different research directions in sound studies.

Stockhausen (1972) further extended the notions of sound materiality and organized sound by taking the discussion to the complex relational dynamics among pitch, noise, timbre and rhythm. In exploring electronic and tape music, Stockhausen framed sound as existing in three-dimensional spaces, where musical elements interact along multiple, often continuous, axes. Later, the temporality of ever-shifting sound properties in composition was further theorized through *spectromorphology* (Smalley, 1997). Together, these new theoretical foundations that emerged in the twentieth century exemplify how once-rigid categories like pitch and noise can be understood as points along a continuum—a principle that strongly resonates with modern machine learning techniques, which situate sounds (or their features) in high-dimensional latent spaces.

Collectively, these mid-century composers prompted a profound shift. Rather than relying solely on discrete tones or canonical instruments, they declared that musical organization could derive from *any* sonic source. This "everything is fair game" ethos naturally converges with the broad ambitions of musical AI, wherein algorithms model, reshape, or generate a diversity of timbres and textures. If Russolo ushered in industrial noise and Varèse rebranded music as "organized sound," machine learning now operationalizes those ideas, providing computational means to *learn* sonic relationships and to *reorganize* audio material at a granular level. By tracing the evolution from Russolo's manifesto to Stockhausen's multi-dimensional conception of sonic space, we see how mid-twentieth-century theories paved the way for an expansive musical canvas. Modern AI frameworks, including neural audio, inherit these conceptual breakthroughs, but also push them further by enabling new levels of autonomy in exploring vast diversity of sounds. This lineage provides essential context for understanding the rise of neural audio instruments, to which we now turn.

# 4 Neural audio instruments

Building on mid-twentieth-century foundations, neural audio instruments represent a pivotal evolution within musical AI. While inheriting the experimental ethos of organized sound and multidimensional sonic exploration, they advance it through machine-learned representations designed for real-time interaction in performance contexts. More specifically, we define neural audio instruments as DMIs that embed neural network and deep learning approaches capable of directly generating or transforming audio signals, and enabling *real-time action–sound mapping* (Jensenius, 2022). Throughout this article, we use the terms *neural audio* and *neural audio models* to refer to such technologies. In what follows, we clarify their distinction from earlier or purely offline AI-driven approaches and explain why they can enable a new class of DMIs.

## 4.1 Neural audio

When embedded in DMIs, neural audio models primarily occupy the *synthesis* side of the design. This departs from earlier, performance-oriented approaches such as *interactive machine learning* (Fiebrink and Cook, 2010), where the model essentially serves as a control conduit, mapping the performer's control inputs to parameters within an arbitrary synthesis engine (e.g., a frequency modulation engine, a sample bank). Those systems employ classification or regression with very small datasets, illustrating how gesture signals should affect synthesis (Fiebrink and Sonami, 2020), thereby maintaining a clear separation between the *control* model and the *synthesis* algorithm.

By contrast, neural audio models learn directly from example audio outputs, embedding within their architectures both the parameters and the algorithms for sound generation. Some rely entirely on neural-network layers (Wright et al., 2020), while others partially integrate neural modules with conventional digital signal processing (DSP) components (Hayes et al., 2024). Here, training data consist of audio waveform samples or higher-level perceptual representations (obtained via spectral transformations like Fast Fourier Transform or Mel-Frequency Cepstral Coefficient extraction), rather than symbolic or gestural inputs. Indeed, entire audio signals can be used to drive the model (Wright et al., 2020; Caillon and Esling, 2021), enabling direct transformations of the incoming sound. In this way, our definition of neural audio also diverges from approaches relying exclusively on symbolic representations (like MIDI) to learn performer–instrument relationships (e.g., Yang et al., 2019).

The result is a family of models that can learn to *organize* or *produce* a broad spectrum of timbres and evolving perceptual phenomena, embracing any and all sounds as legitimate musical material—just as Russolo, Varèse and others once proposed. To achieve this, however, these models frequently demand larger training datasets, capturing enough variety of signals that share the targeted acoustic traits (Caillon and Esling, 2021). Furthemore, these models exhibit *complex relationalities* across modalities. In the visual domain, for example, neural layers can discover progressively higher-level features (e.g., object shape) (Zeiler and Fergus, 2014). Likewise, analogous emergent behaviors arise in the sound domain; Tatar et al. (2020) showed how a Variational Autoencoder (VAE) arranged audio windows from the same recording into nearby regions of latent space, effectively representing sound objects as a continuous "path" in abstract sound possibility spaces.

Crucially, this analysis places neural audio beyond the role of a mere "synthesis module" (like frequency modulation or additive synthesis) within a DMI. Here, the cross-modality and higher-level knowledge imprinted throughout the network manifest in ways that extend beyond sound production. The layers tasked with predicting the next audio sample or shaping the following window are driven by deeper layers that continuously track the historical context of the signal. These deeper layers steer the synthesis process along connections and decisions that hint at what *meaningful* control might look like, e.g., how the network's parameters could/should be modulated to trace a specific sound object's path. This semantic richness informs not only *what* sound can be generated but also *how* it is produced. From this perspective, neural audio instruments depart from the typical independence between control and synthesis observed in canonical DMIs. This stands in stark contrast to interactive machine learning, where users can endlessly tweak mappings between control streams and synthesis parameters, effectively pushing the design process to treat these two elements as conceptually separate. By contrast, leveraging neural audio models suggests a more unified DMI design; one likely characterized by fewer but more intrinsic control options, which are themselves guided by how the model behaves and thus more conducive to an *embodied* sense of playing. Moreover, this influence extends to feedback. The same learned topologies that steer gesture–sound paths can guide tightly-coupled haptic cues, echoing Cadoz's ergodic ideal (Cadoz, 2009) and its role in embodiment. And as models advance toward fully multimodal training (audio-text-vision embeddings, cross-modal VAEs, etc.), DMIs—already intrinsically multisensory—become an ideal platform for embedding and exploring these richer feedback possibilities.

At the same time, the requirement of enabling real-time action–sound mapping, where physical gestures are immediately translated into sonic feedback, raises significant practical constraints. Because these architectures embed learnable synthesis algorithms, they tend to be more computationally intensive than classification- or mapping-oriented networks. Achieving minimal latency for onstage performance thus often necessitates trading off maximal audio fidelity or parameter granularity (Caillon and Esling, 2021). Consequently, while large text-to-music or offline generative models (e.g., Copet et al., 2024; Engel et al., 2019) may exhibit powerful capabilities, they lack the instantaneous feedback loops crucial for DMIs. Neural audio instruments, by contrast, strive for latencies that preserve the tight coupling between performer's gestures and evolving sonic output, fulfilling not only the vision of "organized sound," but also the immediate, corporeal demands of musical embodiment.

## 4.2 Key architectures and their suitability for DMIs

Having established the conceptual underpinnings of neural audio, we now shift to a more technical perspective on how these models are built. The approaches described below reflect distinct strategies for achieving real-time or near-real-time audio generation or transformation, each with particular trade-offs and implications for a performer's action–sound mapping.

### 4.2.1 Sample-by-sample architectures

One early class of neural audio models outputs *one audio sample at a time* via autoregressive techniques. A well-known example is *WaveNet* (Oord et al., 2016a), which builds on PixelCNN (Oord et al., 2016b) by stacking dilated causal convolutions to capture long-range temporal dependencies. Though WaveNet achieved remarkable fidelity–particularly in text-to-speech applications (Oord et al., 2017)–its computational cost poses a hurdle to real-time performance in DMIs. *SampleRNN* (Mehri et al., 2016) similarly generates raw audio sample-by-sample, employing a recurrent neural network trained directly on amplitude values. The architecture can produce 8-bit outputs, as popularized by the music group *dadabots* (Carr and Zukowski, 2018), who created infinite metal and jazz streams.

Despite these compelling use cases, achieving sufficiently low-latency inference for *stage* performance remains an open challenge, limiting widespread adoption of purely sample-level synthesis in DMIs. Autoregressive deep learning approaches as in WaveNET can produce audio sample-by-sample; however, their computational load typically outstrips what conventional setups can handle in real time at high audio rates. Conversely, sample-based approaches have proven quite successful for *audio processing* tasks, where the neural network modifies a continuous input signal *in place*. Indeed, many virtual analog modeling systems leverage sample-by-sample recurrent neural networks for effect emulation (Wright et al., 2020; Hoopes et al., 2024), which can demand fewer complexities than full-blown generative synthesis. In these contexts, the networks primarily learn how to reshape or color the incoming audio stream, requiring less overhead than predicting entirely new waveforms from scratch—thus rendering real-time usage more tractable.

### 4.2.2 Window or spectrogram-based models

A second category of neural audio systems operates at the *audio window* (or spectrogram frame) level, trading sample-level accuracy for computational efficiency. These methods generally follow one of two approaches: directly generating audio windows or producing magnitude spectrograms that are later converted to audio through a vocoder. For example, *RawAudioVAE* (Tatar et al., 2023) encodes short waveform segments (1,024 samples) into a latent space using a VAE. This enables interpolation and manipulation of embedded sonic features, supporting innovative live sound design. *RawAudioVAE* runs ∼5,000× faster than realtime on an NVIDIA RTX 3080-TI graphics card (laptop version). While originally demonstrated in a live-coding context, its low-latency output is suitable for a wider range of DMI designs.

Extending this concept to multimodal applications, Bisig and Tatar (2021) developed *RAMFEM*, which couples a dance-pose encoder with an adversarial VAE for raw audio generation. While this proof-of-concept successfully maps movement trajectories to evolving waveforms, issues with audio fidelity and real-time speed persisted at the time of publication.

Beyond these direct methods, many systems generate magnitude spectrograms (via transforms such as the Short-Time Fourier Transform, the Mel-Frequency Cepstral Transform or the Constant-Q Transform) and then reconstruct the audio signal using vocoders like *Griffin-Lim* (Griffin and Lim, 1984) or *MelGAN* (Kumar et al., 2019). Such two-stage designs allow for conditioning on pitch, loudness, and other descriptors (e.g., Hantrakul et al., 2019; Colonel and Keene, 2020), striking a pragmatic balance between expressive control and manageable computation.

### 4.2.3 Differentiable DSP and sound object generation

A third class integrates DSP blocks within trainable neural networks, focusing on timbral or "sound object" generation while retaining some interpretability. *Differentiable DSP* (Engel et al., 2020) exemplifies this approach by inserting differentiable oscillators, filters and reverberation modules into an end-to-end pipeline, supporting explicit parameter control (pitch, loudness) and timbral transfer across instruments. While differentiable DSP architectures can yield high-quality audio, their control layers often remain somewhat limited in scope (e.g., single parameters for pitch and loudness), constraining them in creative settings.

*Torchsynth* (Turian et al., 2021) similarly merges classic subtractive or additive synthesis modules (oscillators, envelopes, filters) with a graphics card-compatible framework. Its strength lies in generating large datasets for machine learning research or effect design. However, requiring a predefined audio duration before generating sound, it is less suited to fully interactive, real-time engagements; yet still suitable for interactions where the initial sound action is prolonged through processes within the timescale of sound objects, as in the case of a snare hit. Nevertheless, these hybrid neural-DSP pipelines offer a promising avenue for bridging the gap between the richly learned abstractions of deep networks and the user familiarity of standard DSP metaphors.

### 4.2.4 Larger-scale generators

Finally, various deep learning pipelines focus on larger musical timescales, generating entire melodies, beats, or short compositions. Examples include *MusicGen* (Copet et al., 2024), *AudioGen* (Kreuk et al., 2023), and *MusicLM* (Agostinelli et al., 2023), which synthesize musical phrases from textual prompts. As introduced in Section 4.1, these models can produce convincing musical snippets or stylistic transformations, yet they typically lack the real-time feedback loop crucial for DMI performance. Their operational latencies or reliance on batch prompts preclude the immediate, gestural interactions that define neural audio instruments.

## 4.3 Latent spaces and the "organized sound" paradigm

As discussed in Section 3.2, a recurring theme in mid-twentieth-century musical thought was the notion that *all* sounds could be valid compositional material. Composers like Russolo and Varèse envisioned a vast sonic continuum, treating music as "organized sound" rather than discrete pitches or canonical timbres. Neural audio approaches, particularly those involving *latent spaces*, directly extend this legacy by embedding audio data in continuous manifolds. In doing so, they can learn to represent sonic relationships at scale, uncovering structures that mirror and even surpass human perceptual categorizations.

From a technical standpoint, a latent space is a lower-dimensional abstraction learned by models such as VAEs, normalizing flows, or adversarial autoencoders. Rather than relying on hand-engineered parameters or strict symbolic mappings, these networks learn a continuous manifold where acoustically or perceptually similar sounds cluster naturally. In systems like *RawAudioVAE* (Tatar et al., 2023), short audio windows are encoded into latent vectors, potentially giving rise to emergent high-level features (like "timbre" or "noise content") as a consequence of the model's data organization. Yet, latent spaces capture more than isolated similarities; they also reveal the dynamic continuity of sound over time. For instance, as we reported in Section 4.1, consecutive audio windows from the same track tend to form a continuous path in latent space, reflecting transitions that define overall sonic structures (Tatar et al., 2020). Analogous dynamics appear in timbral-transfer systems, where magnitude spectrograms are mapped into latent spaces to discover stylistic similarities (Esling et al., 2018) or to facilitate continuous cross-domain manipulations (Huang et al., 2019). This continuous paths reinforce the notion of *meaningful control* that we previously outlined. Performers navigating these latent trajectories are effectively engaging with the model's internal organization of sound. In other words, the latent space becomes an active interface, one that draws the performer to traverse continuous paths that align with the model's learned notion of "what belongs together" sonically. This deeper understanding of latent spaces elucidates how neural audio systems not only generate sound, but also can serve as a unified approach to real-time, expressive performance.

# 5 Embodiment of neural audio instruments

In analyzing the embodiment of neural audio instruments, it is vital to recognize a longstanding tradition in electronic and digital music wherein *the body* of the sound is effectively severed from its material origins. Early theories in electronic music embraced sound "as is," detached from the objects and events that originally produced it, a legacy that continues to shape our modern view of digital audio data and synthesis. Below, we revisit this historical disassociation, culminating in a renewed understanding of how such perspectives may resonate with or hinder the embodied ethos of neural audio.

Digital audio data and its synthesis are often disassociated from their originating bodies. While in audio synthesis the origin of sound is the electrical currents in analog or digital circuitry, audio recordings have origins in objects, events or processes in real environments. Yet, the audio data, its standards and properties have minimal to no connection to those origins. We think that the causes of disassociation between digital sound data and its originating body lie within the convenience of data storage or

processing in digital and computational means. Tracking the origins of digital audio requires more logistics and labor, given that the affordances of current audio formats and standards do not provide traceability. We argue that this disassociation has resulted in the disembodiment of digital audio and synthesis, as well as digital musical instruments.

In our view, the lack of association between audio in computational domains and embodiment is linked to electronic music history and the phenomenology of sound in early electronic music. Those theories suggested approaching sound without its critical, semantic, functional or connotative properties, and even proposed technological means to *invisibilize* the origins of a sound. Schaeffer (1964) introduced the concept of "objet sonore" or "reduced listening," where the sound is taken purely as sound, without connections to its originating body. Later, John Cage also commented in an interview (Video et al., 2003):

> *When I talk about music, it finally comes to people's minds that I'm talking about sound that doesn't mean anything. That is not inner, but it is just outer. And they say that, these people who understand that finally say, 'you mean it's just sounds?', thinking that for something to just be a sound is to be useless, whereas I love sounds just as they are. And I have no need for them to be anything more than what they are.*

Hence, the phenomenology of sound in early electronic music focused on the materiality of sound disassociated from its body or societal context, connecting to Husserl's notion of *époché* to separate perception from external and cultural factors (Ihde, 2012; Demers, 2010). Later, post-phenomenological criticisms of Schaeffer's "reduced listening" noted that sound exists within a historical and cultural context (Demers, 2010), and listening is more than mere "reduced listening" spanning reflective, denotative, and experiential modes (Tuuri and Eerola, 2012). Specifically, Kane (2007) states two criticisms: "(1) By relying on the sound object to lend an ontological grounding to musical experience, Schaeffer perpetuates an ahistorical view about the nature of musical material [...] (2) Schaeffer maintains an essentialist position on the nature of technology. Rather than re-think the acousmatic reduction in its specific relationship to modern technology, Schaeffer conceives of it as the re-activation of a telos, an originary experience that is presupposed and retained by our practices, yet always available to be re-experienced in its fullness."

In today's audio technologies, we still observe reminiscences of the isolated materiality of sound proposed by both Schaeffer and Cage. Sound is contextualized in computational domains as sound data without its connections to material origins or societal contexts.[8] Audio data is recorded merely as a signal, with no connection to a body. For example, Cotton et al. (2024) presented case scenarios from voice performance and creative speech practices where the disassociation between the recorded voice and its originating body resulted in a loss of control, power and ownership.

Against this backdrop, bridging body and sound in the context of neural audio entails more than just implementing code or collecting data. It may require systematic ways to record and synchronize physical gestures with sonic events, thereby reintroducing "materiality" into neural audio's design and training processes. Ultimately, embodied neural audio may demand new frameworks that capture, imitate, extend, augment or even reinvent the real-world body–sound relationship through computational means. The insights of Schaeffer, Cage and subsequent critics remind us that musical embodiment can be easily severed in a digital paradigm. Before addressing the specific *limitations* and *opportunities* of neural audio instruments (Sections 5.1 and 5.2), or analyzing them from deeper design and philosophical perspectives (Sections 5.3 and 5.4), it is crucial to recognize how historical practices continue to shape the very possibility of an embodied, body-aware approach to digital sound.

## 5.1 Limitations

Neural audio instruments share many of the same embodiment challenges as canonical DMIs, while also introducing additional complexities that can further constrain embodied interaction. One fundamental issue concerns their *symbolic design*. Much like canonical DMIs, the performer's actions are routed through high-level abstractions, but in the case of neural audio instruments these abstractions are considerably more pronounced. The design process involves multiple layers, including model architecture, data curation and training pipelines, each of which can embed a distinct musical or cultural theory into the instrument. In this sense, neural audio instruments realize an even stronger form of what Magnusson (2009) calls "the encapsulation of a specific musical outlook," since their core behaviors derive from symbolic computational structures and dataset-driven knowledge. As Magnusson presciently noted, "[p]articularly in intelligent instruments we find that the expressive design and the determinant of performance experience is to be located at the symbolic computational level," suggesting that AI-based models would inherit and possibly amplify these symbolic constraints. Data curation, in particular, carries strong practical implications, especially when designers must gather the large datasets needed to train neural models. While readily available databases on platforms such as Kaggle[9] and Zenodo[10] can speed up the creation of neural audio instruments, they may also embed cultural norms, biases and stylistic assumptions that limit the performer's scope for personal expression. As noted by Jourdan and Caramiaux (2023), some practitioners resist this "script of technology" and the normative tendencies of big data (see also Cotton and Tatar, 2024), instead favoring smaller, custom datasets that foster a more personal—and potentially more embodied—experience. However, doing so can require substantial time and effort to gather and prepare training materials, prolonging the design process and intensifying the overall challenges of instrument development.

A second challenge derives from the *control bottleneck* (see Section 2), which tends to be even more pronounced in neural audio instruments than in canonical DMIs. Neural

8 An exception to this is the metadata recorded in certain musical audio data, driven partly by copyright or archival needs.

9 https://www.kaggle.com/

10 https://zenodo.org/

models often require larger computational buffers and heavier CPU footprints (Hoopes et al., 2024; Caillon and Esling, 2021), making them less amenable to real-time, low-latency deployment. Consequently, designers might limit the number of mappable parameters to maintain stable performance, thus narrowing the sensorimotor bandwidth between performer and instrument (Privato et al., 2024). Unsurprisingly, some practitioners prefer to avoid the heavy deep learning architectures that power neural audio DMIs. Instead, they embed simpler machine learning algorithms (e.g., regression models) on the *control* side of the instrument rather than in *synthesis*, preserving the liveliness and spontaneity crucial for musical expression (Jourdan and Caramiaux, 2023). Magnusson (2009) stresses that "human actions are displaced into representation, thus establishing strata of complexities and interdependencies that limit the agent"; in neural audio instruments, these strata are deepened by learning-based architectures, latent spaces and opaque model states. The result can be a heightened sense of detachment, especially when performers cannot easily discern which parameters are being controlled or how changes in control gestures translate to audible outcomes. While symbolic interfaces need not disrupt flow *in principle* (Section 2), in practice the limited resolution and the unpredictable jitter of AI-driven systems often lead to the disruption in flow that Magnusson attributes to symbolic music systems.

Building on these phenomenological complexities, neural architectures can also exacerbate the *black box effect*, posing not only a challenge to the performer's sense of embodiment but also raising *epistemological* concerns. In most DMIs, the software and hardware processes that define the relationship between action and sound are inaccessible to the performer, who consequently perceives the instrument as an immutable, sealed technology (i.e., a black box) (McPherson and Zappi, 2015). Even when the effect of musical actions is obvious from the outside, this inaccessibility makes it significantly harder to explore, tweak and appropriate the instrument. The hurdle becomes even more prominent when DMIs embed neural audio and latent spaces, where the model's learned sound organization and latent vector ontology further obscure how actions translate to audible results, often leading to distrust or frustration (Jourdan and Caramiaux, 2023). Because neural models often operate on hidden layers of learned representations, it can be challenging for musicians to develop a deep, embodied understanding of the instrument's behavior. While we argue that the lack of explainability is not always a hindrance to embodiment (see next subsection), this additional mediating layer can complicate performance and prevent the instrument from being fully internalized as an extension of the musician's body and intentions.

Importantly, we argue that disruptions in flow and embodiment are not merely a consequence of enhanced *symbolic* design, large-scale machine-learning techniques, or high computational demand; they are also attributable to inadequate control paradigms. In other words, much like canonical DMIs, a significant part of the responsibility lies with designers and the choices they make (Gurevich, 2014). In the case of neural audio instruments, the continuous mode switches between the DMI and the "terminus of our activities" (i.e., the musical outcome) described by Magnusson (2009) risk being exacerbated by mismatches between the performer's intention and the system's learned behavior. These mismatches—deriving by both the training process and the underlying technology—can break concentration and scatter the performer's focus, the precise antithesis of fluid, embodied engagement. While certain specificities of neural audio may require unique countermeasures (more on this in the next section), this scenario underscores a heightened need for designers to experiment with mapping and real-time control solutions that align with the vast body of research on DMI embodiment (see Section 2.2). Otherwise, the risk is to amplify the very problems of abstraction and detachment that have historically affected DMIs in comparison to acoustic instruments.

## 5.2 Opportunities

Although inheriting design paradigms from canonical DMIs, neural audio instruments can foster an embodiment reminiscent of acoustic instruments. This is primarily due to the *lack of explainability* in their inner workings and the fundamentally different nature of their design process. Unlike imperative audio coding, where developers must possess a "solid theoretical knowledge of sound" (Magnusson, 2009), trained neural audio networks often *include* knowledge that the designers themselves may not fully grasp—mirroring how the crafting of acoustic instruments may not necessitate complete scientific mastery of the equations and models governing the physics of materials and vibrations. This built-in opacity can, paradoxically, encourage *non-theoretical knowledge* rooted in "discovery, exploration, and refinement," reinforcing embodiment through trial and error (Magnusson, 2009). In line with this, neural audio instruments lack the prescribed schematics and instruction manuals that typically guide the use of analog or digital electronic instruments. While technical documentation may exist, there is no immutable description of latent spaces or straightforward interpretation of learned parameters (Tatar et al., 2020; Privato et al., 2024), reinforcing the parallel with acoustic instruments and the process of gradually discovering their quirks and their sonic affordances.

This relative independence from explicit design theory is magnified by the often larger gap between designer and performer in neural audio contexts. While canonical DMIs commonly involve overlapping roles or close collaboration (Morreale et al., 2018), many neural audio models are now created by AI specialists and only later appropriated by musicians (Chowdhury, 2021; Jourdan and Caramiaux, 2023), whose expertise leans more toward hands-on artistic practice than toward scientific or technical knowledge of the model's architecture. This division of labor can spare neural audio practitioners from the tension between conceptual design and embodied performance that DMI musicians must often reconcile in their practice (Magnusson, 2009). Rather than contending with every layer of code that formalizes the relationship between model architecture, datasets and behavior at inference time, these performers can learn neural audio instruments *in situ*, coaxing out unexpected sonic behaviors through embodied experimentation, much as acoustic musicians internalize the tactile features of their instruments. And in the context of this type of

exploration, the lines that divide intentional affordances, hidden modes of interaction and constraints may become increasingly blurred, inevitably shaping how distinctive performance styles and personal playing techniques take form (Zappi and McPherson, 2014). In other words, the complex and often opaque interplay of model architecture and data may hinder performers' ability to discern explicitly designed behaviors from emerging ones, creating a sense of ambiguity that can encourage discovery and physical appropriation (Masu et al., 2016). And the familiar feeling of playing a vastly appropriated instrument is quite likely to serve as a catalyst for embodied interaction.

A further consequence of this larger gap between designers and performers emerges once the neural model has been trained. While musicians can engage in in-depth exploration of a trained network, it is far more difficult to go back and refine its behavior or control interface than it is in most canonical DMIs. In the latter, revising mappings between synthesis and control (and feedback too) is often an integral part of an instrument's ongoing exploration and practice, rather than a task restricted to the initial design. Caramiaux et al. (2014) emphasize that postponing certain design choices to performers can be essential, "allowing them to interactively implement their own metaphors and control strategies." This fluid approach to mapping even led Laetitia Sonami to declare during her ICMC 2024 keynote that "the mapping *is* the instrument" (Sonami, 2024).

Beyond mapping, neural audio instruments complicate the whole concept of fluid or iterative design. Introducing a new conditioning parameter may demand a complete update of the model architecture, a fresh dataset and a second (often time-consuming) training cycle. Practitioners liken this limitation to working with analog electronic instruments—design possibilities may be broad, but once committed to a particular configuration, reversing course can prove infeasible. The computational burden can be daunting as well. Recent advances in deep learning often require resources beyond a single artist's reach (Jourdan and Caramiaux, 2023) and as Sonami noted during her keynote, "it takes three weeks; it's so frustrating," or "I could not [train and test] [...] on my computer," leaving creators feeling as though they possess "powerful tools that we cannot use." While fine-tuning or transfer learning might mitigate these hurdles, many practitioners (Sonami included) remain reluctant to spend hours recording new datasets or re-engineering architectures.

In this sense, instrument-making in neural audio contexts begins to resemble acoustic instrument-making, like luthiery. Luthiers rarely deviate drastically from a starting design, since significant material changes may kill the very resonance they aim to preserve. Similarly, major revisions to a trained network or a neural audio architecture can undermine carefully tuned behaviors, forcing one effectively to rebuild the instrument from scratch. In canonical DMIs, by contrast, large changes beyond "safe" remappings are typically easier to implement—adding filters or output channels, swapping synthesis techniques and so on. As a result, musicians working with neural audio instruments are often driven to explore the instrument *as-is*, identifying and appropriating its innate affordances rather than perpetually retooling its internal structure. This exploratory mindset can encourage a path to embodiment uninhibited by theoretical knowledge of model design, again mirroring the organic processes of acoustic instrument practice. Nevertheless, the constraints described so far need not entirely stifle the impulse to reimagine and appropriate the instrument. Rather, they can channel it in ways reminiscent of physically preparing a piano, as when musicians introduce external interventions (e.g., added materials or tangible scores, as shown by Privato et al., 2024) that do not alter a model's core, but instead reshape the instrument's expressive scope without revolutionizing its inner structure and mechanisms.

The same long retraining cycle that makes radical redesigns impractical also imposes a fruitful form of *slowness* that is directly aligned with recent perspectives in human-computer interaction (Odom, 2024). Because neural-audio runs can occupy anything from half a day to several weeks, design exploration and aesthetic experiments unfold far more slowly than in non-autonomous software, where code changes are heard at once. This enforced pace compels designers to engage with the technology reflectively, rather than through rapid modifications based on trial and error. Such slowness in neural audio development emphasizes the role of artistic experimentation and, in particular, of critical listening (Demers, 2010; Tuuri and Eerola, 2012). While loss curves and other metrics offer some guidance, the sonic qualities of a trained model matter most in a musical setting and should be judged through careful human or machine listening. Designers therefore cycle between audition, small dataset or hyper-parameter tweaks and another long training run, learning the instrument's behavior *in situ* through sound, rather than via rapid code edits. This rhythm restores listening to the center of instrument making and, by anchoring exploration in the aural domain, promises to reinforce the performer's embodied connection to the emerging tool.

## 5.3 Where design meets philosophy

A more interdisciplinary analysis of neural audio instruments reveals subtler and sometimes less straightforward aspects of embodiment, emerging from the intersection of design, human-computer interaction and philosophy. Unlike many established music technologies that are largely taken for granted, neural audio systems frequently undergo heightened scrutiny. As Laetitia Sonami reflected during her ICMC 2024 keynote, "can't [the neural audio model] create something new, something great? [...] It sounds like a failed rendition of the original result." A similar skepticism can be found among other artists and practitioners (Jourdan and Caramiaux, 2023), suggesting that the *concretisation* (Latour, 1987) or widespread acceptance of neural audio tools is not yet in full swing. In other words, despite the rapid ascent of machine learning in industry, academia and art, neural audio has yet to be seamlessly blackboxed[11] into a standardized practice. Two converging factors appear to underlie this provisional status: first, the technology is still maturing and has not yet attained the simplicity or reliability necessary for broad,

11 Not to be confused with the "black box effect" in DMIs, where the *design* is opaque to the performer. Here, "blackboxing" refers to Latour's concept of scientific or technological acceptance.

frictionless adoption; and second, the artistic realm is inherently more critical, where creators often ask whether a tool truly expands the expressive palette or merely replicates existing paradigms, even when the general audience is fully convinced of the tool's quality/effectiveness. In turn, this ambivalence can constrain full-bodied immersion; it is indeed unclear how deeply performers can embody an instrument whose output they distrust or regard as not quite "there" yet.

A distinct layer of insight emerges when examining what Magnusson (2009) calls the *hermeneutic* quality of digital instruments. Neural audio instruments, like other DMIs, exist as a "medium for a hermeneutic relation," but their reliance on training data means that performers are not only interpreting the instrument. They are also, at some stage, teaching it to interpret and modify the world. Once and if purposefully trained, the instrument externalizes a large portion of the musician's aesthetics, extending both their mind and their body. Musicians then find themselves operating "in sync" with a device that exhibits unpredictable, nonlinear behaviors akin to acoustic instruments, yet in a higher-level, learned domain. This dynamic resonates with Sonami's discussion of "unpredictable" machine learning, where a broad training set can produce large, unexpected movements in "timbre space," whereas a narrower set yields more stable oscillations. Such flexibility allows the performer to scale or recalibrate the "predictability index" (Fiebrink and Sonami, 2020), offering novel opportunities for real-time musical adaptation and go beyond hermeneutics.

From a *soma* perspective, this peculiar extension of human musical capabilities through neural audio instruments disrupts traditional dualisms between the performer's physical self and the surrounding material environment. In the language of Höök et al. (2021), a symbiosis emerges in which human and machine take turns exerting control over the musical process. Although the model now takes on part of the performer's creative burden (in some sort of cognitive delegation), continuous bodily engagement remains essential for guiding the instrument's responses in real time. Moreover, the machine is not completely cognitively independent either, inasmuch as the musician's aesthetic choices and artistic strategies have already been "taught" to the model through training data, imbuing the instrument with a partial manifestation of human intent. By integrating the unpredictability and learned capacities of AI into the musician's sensorimotor loop, these systems can foster an unprecedented form of *integrated embodiment* that transcends simple augmentation (Hu et al., 2017), weaving mind, body and machine into one evolving musical agent. This vision contrasts with Magnusson, who describes a *post-human intentionality* emerging in interactions with instruments that exhibit agency [e.g., feedback systems (Magnusson et al., 2022), AI-driven DMIs]. Drawing on Ihde's concept of *alterity* (Ihde, 1990), Magnusson argues that the AI instrument represents an "otherness," i.e., something that "is not an extension of us or our thinking" (Magnusson, 2023). In contrast, our framework leans toward what we call a *trans-human intentionality*, manifested through integrated embodiment with neural audio technologies. Rather than emphasizing the instrument as an autonomous other, we envision a co-evolving partnership beyond physical and cognitive dualisms.

## 5.4 Musical agency

Since the outset of this article, we have emphasized that our interpretation of neural audio departs from the concept of a *musical agent*. The design of neural audio models embedded in DMIs, as discussed in Section 4.1, does not primarily address the challenges of creating an improvising collaborator that actively responds to a musician's performance. Rather, we have framed neural audio architectures as an advanced generation of sound-producing modules that can be controlled in real time, shaping an *instrumental* relationship between system and performer. Yet, because these networks inherently encode control semantics (e.g., latent spaces, learned representations), they also steer and inspire the performer's agency in ways canonical DMI modules generally do not.

Nevertheless, it is not uncommon for players to experience neural audio models in agent-like terms. For instance, Laetitia Sonami, reflecting on her music-making experiences with neural audio instruments at the 2024 ICMC keynote (Sonami, 2024), remarked: "I feel like having a conversation," while other accounts offer equally direct expressions of agency (Erdem, 2022). Rather than being contradictory, these sentiments point to a deeper interpretive framework wherein partially unpredictable, learned behaviors elicit a sense of autonomy or co-performance. Drawing on Moore (2024), several contingencies can foster a performer's perception of agency, many of which can resonate with neural audio: (i) *analog contingency*, where a musician cannot reliably replicate an instrument's sonic output; (ii) *improvisational contingency*, in which the system's unpredictable responses prompt spontaneous, adaptive playing; and (iii) *time-varying contingency*, whereby mid- and large-scale consistency coexists with momentary surprises. Additionally, Moore notes that agency may emerge whenever system complexity foils the user's ability to track every parameter—precisely the kind of "black box" situation often posed by learned parameters and latent spaces in neural audio.

An apparently corollary detail, latency, also shapes whether performers interpret the instrument as an agent. Lengthy response times can make the system feel more like a collaborator with its own timeframe than a responsive tool. Anecdotal and documented experiences from musicians[12] underscore that excessive latency can foster the impression of the instrument acting on its own, rather than simply reacting to the performer's commands. Similar remarks by Sonami in her keynote emphasize how "repetition" and "echo" cycles lead to disjointed interactions that some interpret as otherness or an agent-like presence (Magnusson, 2023).

However, Moore also notes that true collaborative agents are typically *intended* and *perceived* as agential from the outset. If a neural audio system's unexpected output is primarily seen as an error or glitch—something designers strive to eliminate, rather than an intentional creative deviation—the relationship aligns less with agency and more with the constraints

12 *inSonic 2020: syntheses* Festival, 12 December 2020, Karlsruhe, Germany. Recording available at https://www.youtube.com/watch?v=sooNxK6oQ4c.

and idiosyncrasies of an instrument. Indeed, current neural audio research often focuses on improving predictability, responsiveness and quality (Shier et al., 2024), reinforcing the view that these models, while sometimes interpreted as having their own "will," are ultimately created to be instrumental to interactive design.

Finally, Moore (2024) also observes that many practitioners link a system's *intelligence* directly to the "perception or attribution of agency." From this perspective, the question of how "intelligent" neural audio models truly are remains ambiguous. Some accounts of intelligence, particularly in artificial contexts, prioritize learning as the core indicator of cognition, providing a more favorable framing for neural audio. However, if we adopt the *situated* view articulated by Suchman (1987), intelligence emerges through ongoing, context-sensitive adaptation. In neural audio models, this adaptability remains largely confined to the training phase, "freezing" network parameters before performance. Consequently, there is minimal scope for the instrument to dynamically respond to environmental changes. Indeed, even the presence of genuine cognition itself may be questioned (Erdem, 2022), since the model lacks both a sense of environment and a mechanism to continually refine its actions *in situ*. In a situated perspective on intelligence, the agent's cognition emerges from a dynamic interplay between bodily engagement, sensorimotor feedback and an evolving environment (Suchman, 1987). By contrast, AI models' (including both neural audio models and musical agents) frozen parameters leave little room for adaptive re-learning, or context-sensitive adaptation. Discarding the training–inference dichotomy and enabling continuous learning during performance could, in principle, ground a more fully embodied or agentic neural audio practice, with the model actively sensing its environment and reshaping its internal representations in real time. For now, however, whether one perceives these instruments as "intelligent" or merely "instrumental" often depends on lived experience, rather than any universal criterion.

# 6 Conclusion

In this article, we set out to examine how the concept of musical embodiment intersects with the emerging domain of *neural audio* instrument design. We began by contextualizing embodiment through literature on traditional and acoustic instruments, and then extended the focus to DMIs. Against this backdrop, we introduced the idea of neural audio instruments within a wider landscape of AI-driven music technologies, discussing both their mid-twentieth-century aesthetic roots and the various neural architectures that underpin them. These explorations provided a foundation for understanding their potential within embodied musical practices. From there, we considered how such instruments might be *embodied* in practice. We highlighted the ways in which they inherit the limitations of DMIs, yet also open up novel possibilities thanks to their unique technological paradigms. By integrating perspectives from design, human-computer interaction and philosophy, we highlighted both the potential stumbling blocks and surprising opportunities for creating deeply expressive musical encounters with AI-driven instruments.

Throughout this analysis, we encountered apparent contradictions. For example, on one hand, neural audio instruments often challenge performers by resisting traditional approaches to remapping or modifying inner workings; their opaque architectures and learned representations can be difficult to grasp or adapt. On the other hand, this very opacity can liberate musicians from technological pre-conceptions or a strictly hermeneutic mindset, allowing them to engage with the instrument through tacit exploration rather than forming a rigid theoretical model. For designers, the realization of this and similar contradictions might seem daunting; they underscore how every nuanced design decision (and not just those tied to technical feasibility) can reverberate through the musician's subjective, embodied relationship with this specific manifestation of musical AI. Still, in the spirit of Gurevich (2014), who concluded his own challenging assessment of skill in DMIs with tempered optimism, our intention "is not to say that all hope is lost." As our discussion emphasizes, new design initiatives continue to emerge at the intersection of technology, music, science and philosophy. Together with the growing ecosystem of neural audio toolkits and creative communities, such endeavors offer promising, even disruptive paradigms for future instrument makers willing to embrace complexity and experiment with the aesthetics and affordances of AI.

Along these lines, we would like to conclude with five *practical considerations* on how to design neural audio instruments that come from what discussed so far and are aimed at fostering a strong sense of embodiment. We hope such guidelines will support designers that today are at the forefront of this exciting yet challenging endeavor, by encouraging them to create instruments that empower deeply musical and profoundly human experiences in partnership with learning-based technologies.

**0. Stand on the shoulders of giants.** All the core insights from the DMI literature (see Section 2) remain relevant—and perhaps even more critical—when designing neural audio instruments. Challenges like the control bottleneck and the symbolic nature of action-to-sound can become more pronounced under AI-driven conditions, so established guidance on fostering embodiment in DMIs still applies here as a vital starting point!

**1. Search for new modes of interaction.** As Magnusson (2009) notes, the behaviors and "materials" of any instrument strongly condition how musicians interact with it. Neural networks, however, may exhibit properties not easily paralleled in earlier instruments. Hence, novel paradigms, such as directly "traversing" multi-dimensional latent spaces, might offer fresh avenues for mapping movement and cognition to sonic outcomes, potentially unlocking more intuitive or embodied interfaces than one might initially assume.

**2. Challenge dualities.** Somaesthetics and phenomenology already question dualities like mind–body and body–environment (Höök et al., 2021) and future work on neural audio may similarly challenge a strict training–inference split (Section 5.4). Moreover, a pressing and practical concern for DMIs lies in the traditional *control–synthesis* divide and the predicate of mapping. We do not suggest abandoning mapping altogether; exploration of how

gesture connects to sound is a valuable design tool. However, we advocate a holistic design perspective where sound and gesture are conceived as a unified entity from the outset (Caramiaux et al., 2014), rather than as two separate "containers" later bound by mapping. Once this integrated foundation has been laid, the technological challenges of AI, neural networks, data and training can be addressed without losing an inherently embodied connection between motion and sonic outcome. This approach paves the way for more inventive metaphors and interaction techniques that surpass mere iterative adjustments to input and output streams.

**3. Embrace inexplicability (with a grain of salt).** While research on explainable AI is undoubtedly worthwhile, non-explainability can play a significant role in the use and design of neural audio instruments. A flute player, for instance, need not fully grasp the acoustic physics behind overtone production to exploit them masterfully. Likewise, performers and even designers of neural audio systems may choose to focus on *musical outcomes* rather than dissecting every underlying process. Indeed, not all instrument designs are "predicated on the application of scientific knowledge" (Green, 2011) and a certain measure of "unknowing" can inspire extraordinary results. By positioning neural audio synthesis at the intersection of scientific modeling and pure intuition, designers can open pathways to creative strategies unattainable through rational design alone. This notion also resonates with broader human-computer interaction discourse on the creative power of *ignorance* (Grammenos, 2014) (ranging from lack of preconceptions, to *true* ignorance), where "if you already know where you are going, you are not going someplace new."

**4. Make AI inconspicuous.** When the AI is not intended to act as a distinct musical *agent*, making its presence explicit may be unnecessary or even counterproductive. Instead, designers might treat neural audio models as just another invisible part of the instrument's anatomy, like the string of a piano or the integrated circuit of an analog synthesizer. By letting the model manifest itself only through the *embodiment* of the musician's actions and intentions (the trans-human intentionality), the performer can experience a unified instrument rather than a model endowed with conspicuous (artificial) intelligence. Under the hood, such intelligence may enable feats that would otherwise be impossible, such as large-scale physical modeling (Diaz et al., 2023), multi-stream data handling (Fiebrink and Sonami, 2020), or high-level perceptual organization (Tatar et al., 2020). Yet performers need not be confronted with "AI" *per se*. By rendering the model seamlessly integral, designers promote an experience of *playing* an instrument rather than *interfacing* with an AI model.

## Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

## Author contributions

VZ: Conceptualization, Data curation, Methodology, Resources, Supervision, Writing – original draft, Writing – review & editing. KT: Conceptualization, Data curation, Funding acquisition, Methodology, Resources, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program—Humanity and Society (WASP-HS), funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Gen AI is used for text proofreading and editing of this article.

## Publisher's note



## References

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., et al. (2023). Musiclm: generating music from text. *arXiv [preprint]* arXiv:2301.11325.

Armstrong, N. (2006). *An Enactive Approach to Digital Musical Instrument Design* [PhD thesis]. Princeton University, Princeton, NJ.

Arte Video, Société des auteurs, compositeurs et éditeurs de musique, Centre Georges Pompidou, Centre national de la cinématographie, ARTE France, et al. (2003). *Listen = [Ecoute] [Video recording]*. Issy-les-Moulineaux: ARTE France Développement.

Bang, T. G., and Fdili Alaoui, S. (2023). "Suspended circles: soma designing a musical instrument," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–15. doi: 10.1145/3544548.3581488

Begleiter, R., El-Yaniv, R., and Yona, G. (2004). On prediction using variable order Markov models. *J. Artif. Intell. Res.* 22, 385–421. doi: 10.1613/jair.1491

Berthaut, F., Desainte-Catherine, M., and Hachet, M. (2011). Interacting with 3D reactive widgets for musical performance.

*J. New Music Res.* 40, 253–263. doi: 10.1080/09298215.2011.602693

Bisig, D., and Tatar, K. (2021). "Raw music from free movements: early experiments in using machine learning to create raw audio from dance movements," in *Proceedings of the 2nd AI Music Creativity Conference*, 11.

Briot, J.-P., and Pachet, F. (2020). Deep learning for music generation: challenges and directions. *Neural Comput. Appl.* 32, 981–993. doi: 10.1007/s00521-018-3813-6

Brown, A. R., and Sorensen, A. (2009). Interacting with generative music through live coding. *Contemp. Music Rev.* 28, 17–29. doi: 10.1080/07494460802663991

Burland, K., and McLean, A. (2016). Understanding live coding events. *Int. J. Perform. Arts Digit. Media* 12, 139–151. doi: 10.1080/14794713.2016.1227596

Cadoz, C. (2009). Supra-instrumental interactions and gestures. *J. New Music Res.* 38, 215–230. doi: 10.1080/09298210903137641

Caillon, A., and Esling, P. (2021). Rave: a variational autoencoder for fast and high-quality neural audio synthesis. *arXiv* [Preprint]. arXiv:2111.05011. doi: 10.48550/arXiv.2111.05011

Çamcı, A., and Granzow, J. (2019). Hyperreal instruments: bridging VR and digital fabrication to facilitate new forms of musical expression. *Leonardo Music J.* 29, 14–18. doi: 10.1162/lmj_a_01056

Cannon, J., and Favilla, S. (2012). "The investment of play: expression and affordances in digital musical instrument design," in *Proceedings of the International Computer Music Conference*, 459–456.

Caramiaux, B., Donnarumma, M., and Tanaka, A. (2015). Understanding gesture expressivity through muscle sensing. *ACM Trans. Comput.-Hum. Interact.* 21, 1–26. doi: 10.1145/2687922

Caramiaux, B., Françoise, J., Schnell, N., and Bevilacqua, F. (2014). Mapping through listening. *Comput. Music J.* 38, 34–48. doi: 10.1162/COMJ_a_00255

Carnovalini, F., and Rodà, A. (2020). Computational creativity and music generation systems: an introduction to the state of the art. *Front. Artif. Intell.* 3:14. doi: 10.3389/frai.2020.00014

Carr, C. J., and Zukowski, Z. (2018). Generating albums with sampleRNN to imitate metal, rock, and punk bands. *arXiv* [Preprint] arXiv:1811.06633. doi: 10.48550/arXiv.1811.06633

Cavdir, D., and Dahl, S. (2022). "Performers' use of space and body in movement interaction with a movement-based digital musical instrument," in *Proceedings of the 8th International Conference on Movement and Computing* (New York, NY: ACM), 1–12. doi: 10.1145/3537972.3537976

Chirico, A., Serino, S., Cipresso, P., Gaggioli, A., and Riva, G. (2015). When music "flows". state and trait in musical performance, composition and listening: a systematic review. *Front. Psychol.* 6:906. doi: 10.3389/fpsyg.2015.00906

Chowdhury, J. (2021). Rtneural: fast neural inferencing for real-time systems. *arXiv* [Preprint] arXiv:2106.03037. doi: 10.48550/arXiv.2106.03037

Colonel, J. T., and Keene, S. (2020). "Conditioning autoencoder latent spaces for real-time timbre interpolation and synthesis," in *2020 International Joint Conference on Neural Networks (IJCNN)* (Glasgow: IEEE), 1–7. doi: 10.1109/IJCNN48605.2020.9207666

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., et al. (2024). "Simple and controllable music generation," in *Advances in Neural Information Processing Systems 36*.

Cotton, K., de Vries, K., and Tatar, K. (2024). "Singing for the missing: bringing the body back to ai voice and speech technologies," in *Proceedings of the 9th International Conference on Movement and Computing* (New York, NY: Association for Computing Machinery). doi: 10.1145/3658852.3659065

Cotton, K., and Tatar, K. (2024). "Sounding out extra-normal AI voice: non-normative musical engagements with normative AI voice and speech technologies," in *Proceedings of the 5th AI Music Creativity Conference*. Available online at: https://aimc2024.pubpub.org/pub/extranormal-aivoice (Accessed August 1, 2025).

Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety*. San Francisco, CA: Jossey-bass.

Dahl, L. (2014). "Triggering sounds from discrete air gestures: what movement feature has the best timing?" in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 201–206.

Davison, M., Webb, C. J., Ducceschi, M., McPherson, A. P., et al. (2024). "A self-sensing haptic actuator for tactile interaction with physical modelling synthesis," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 574–581.

de Lima Costa, W., Filgueira, D., Ananias, L., Barioni, R., Figueiredo, L. S., and Teichrieb, V. (2020). Songverse: a digital musical instrument based on virtual reality. *J. Interact. Syst.* 11, 57–65. doi: 10.5753/jis.2020.749

Demers, J. T. (2010). *Listening Through the Noise: The Aesthetics of Experimental Electronic Music*. Oxford: Oxford University Press.

Diaz, R., Saitis, C., and Sandler, M. (2023). "Interactive neural resonators," in *Proceedings of the International Conference on New Interfaces for Musical Expression*. PubPub.

Dourish, P. (2001). *Where the Action is: The Foundations of Embodied Interaction, volume 210*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/7221.001.0001

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., Roberts, A., et al. (2019). "Gansynth: adversarial neural audio synthesis," in *Proceedings of The Seventh International Conference on Learning Representations*.

Engel, J., Hantrakul, L. H., Gu, C., and Roberts, A. (2020). "DDSP: differentiable digital signal processing," in *Proceedings of The 8th International Conference on Learning Representations*.

Erdem, Ç. (2022). "Exploring musical agents with embodied perspectives," in *International Seminar on Sonic Design* (Cham: Springer), 321–341. doi: 10.1007/978-3-031-57892-2_17

Esling, P., Chemla-Romeu-Santos, A., and Bitton, A. (2018). "Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics," in *Proceedings of the 21st International Conference on Digital Audio Effects* (Aveiro), 369–366.

Essl, G., and O'Modhrain, S. (2006). An enactive approach to the design of new tangible musical instruments. *Organised Sound* 11, 285–296. doi: 10.1017/S135577180600152X

Fels, S. (2004). Designing for intimacy: creating new interfaces for musical expression. *Proc. IEEE* 92, 672–685. doi: 10.1109/JPROC.2004.825887

Fernández Rodríguez, J. D., and Vico Vela, F. J. (2013). AI methods in algorithmic composition: a comprehensive survey. *J. Artif. Intell. Res.* 48, 513–582. doi: 10.1613/jair.3908

Fiebrink, R., and Cook, P. R. (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, Vol. 3* (Utrecht: Citeseer), 2–1.

Fiebrink, R., and Sonami, L. (2020). "Reflections on eight years of instrument creation with machine learning," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 237–242.

Fowler, C. B. (1967). The museum of music: a history of mechanical instruments. *Music Educ. J.* 54:45. doi: 10.2307/3391092

Galanter, P. (2003). "What is generative art? Complexity theory as a context for art theory," in *Proceedings of the 6th Generative Art Conference*.

Godøy, R. I. (2017). "Postures and motion shaping musical experience," in *The Routledge Companion to Embodied Music Interaction* (London: Routledge), 113–121. doi: 10.4324/9781315621364-13

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Grammenos, D. (2014). "Abba-dabba-ooga-booga-hoojee-goojee-yabba-dabba-doo: stupidity, ignorance & nonsense as tools for nurturing creative thinking," in *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (New York, NY: ACM), 695–706. doi: 10.1145/2559206.2578860

Green, O. (2011). Agility and playfulness: technology and skill in the performance ecosystem. *Organised Sound* 16, 134–144. doi: 10.1017/S1355771811000082

Griffin, D. W., and Lim, J. S. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust.* 32, 236–243. doi: 10.1109/TASSP.1984.1164317

Gurevich, M. (2014). "Skill in interactive digital music systems," in *The Oxford Handbook of Interactive Audio*, eds. K. Collins, B. Kapralos, and H. Tessler (Cham: Springer), 315–332.

Gurevich, M., and Fyans, A. C. (2011). Digital musical interactions: performer-system relationships and their perception by spectators. *Organised Sound* 16, 166–175. doi: 10.1017/S1355771811000112

Habe, K., Biasutti, M., and Kajtna, T. (2019). Flow and satisfaction with life in elite musicians and top athletes. *Front. Psychol.* 10:698. doi: 10.3389/fpsyg.2019.00698

Hantrakul, L., Engel, J., Roberts, A., and Gu, C. (2019). "Fast and flexible neural audio synthesis," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*, 7.

Hayes, B., Shier, J., Fazekas, G., McPherson, A., and Saitis, C. (2024). A review of differentiable digital signal processing for music and speech synthesis. *Front. Signal Process.* 3:1284100. doi: 10.3389/frsip.2023.1284100

Hayes, L. (2022). "Why should we care about the body?: On what enactive-ecological musical approaches have to offer," in *The Body in Sound, Music and Performance* (Waltham, MA: Focal Press), 23–34. doi: 10.4324/9781003008217-4

Herremans, D., Chuan, C.-H., and Chew, E. (2017). A functional taxonomy of music generation systems. *ACM Comput. Surv.* 50, 1–30. doi: 10.1145/3108242

Höök, K. (2018). *Designing with the Body: Somaesthetic Interaction Design*. Cambridge, MA: MIT Press.

Höök, K., Benford, S., Tennent, P., Tsaknaki, V., Alfaras, M., Avila, J. M., et al. (2021). Unpacking non-dualistic design: the soma design case. *ACM Trans. Comput.-Hum. Interact.* 28, 1–36. doi: 10.1145/3462448

Hoopes, J., Chalmers, B., and Zappi, V. (2024). "Neural audio processing on android phones," in *Proceedings of the 27th International Conference on Digital Audio Effects*, 33–40.

Hu, Y. Leigh, S.-w., Maes, P. (2017). "Hand development kit: soft robotic fingers as prosthetic augmentation of the hand," in *Adjunct Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (New York, NY: ACM), 27–29. doi: 10.1145/3131785.3131805

Huang, S., Li, Q., Anil, C., Bao, X., Oore, S., Grosse, R. B., et al. (2019). "TimbreTron: a WaveNet(CycleGAN(CQT(Audio))) pipeline for musical timbre transfer," in *Proceedings of The 7th International Conference on Learning Representations*. doi: 10.48550/arXiv.1811.09620

Hunt, A., and Wanderley, M. (2002). Mapping performance parameters to synthesis engine. *Organised Sound* 7, 103–114. doi: 10.1017/S1355771802002030

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Bloomington; Indianapolis, IN: Indiana University Press.

Ihde, D. (2012). *Listening and Voice: Phenomenologies of Sound*. New York, NY: State University of New York Press.

Ivanyi, B. A., Tjemsland, T. B., Tsalidis de Zabala, C. V., Toth, L. J., Dyrholm, M. A., Naylor, S. J., et al. (2023). "Duorhythmo: design and remote user experience evaluation (UXE) of a collaborative accessible digital musical interface (CADMI) for people with ALS (PALS)," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–13. doi: 10.1145/3544548.3581285

Jack, R. H., Stockman, T., and McPherson, A. (2017a). "Maintaining and constraining performer touch in the design of digital musical instruments," in *Proceedings of the Eleventh International Conference on Tangible, Embedded, and Embodied Interaction* (New York, NY: ACM), 717–720. doi: 10.1145/3024969.3025042

Jack, R. H., Stockman, T., and McPherson, A. (2017b). "Rich gesture, reduced control: the influence of constrained mappings on performance technique," in *Proceedings of the 4th International Conference on Movement Computing* (New York, NY: ACM), 1–8. doi: 10.1145/3077981.3078039

Jackson, S. A., and Eklund, R. C. (2002). Assessing flow in physical activity: the flow state scale-2 and dispositional flow scale-2. *J. Sport Exerc. Psychol.* 24, 133–150. doi: 10.1123/jsep.24.2.133

Jensenius, A. R. (2014). "To gesture or not? an analysis of terminology in NIME proceedings 2001-2013," in *A NIME Reader: Fifteen Years of New Interfaces for Musical Expression* (Cham: Springer), 451–464. doi: 10.1007/978-3-319-47214-0_29

Jensenius, A. R. (2022). *Sound Actions: Conceptualizing Musical Instruments*. Cambridge: MIT Press. doi: 10.7551/mitpress/14220.001.0001

Jensenius, A. R., and Wanderley, M. M. (2010). "Musical gestures: concepts and methods in research," in *Musical Gestures* (London: Routledge), 24–47.

Jordà, S. (2004). "Digital instruments and players: part I–efficiency and apprenticeship," in *Proceedings of the International Conference on New interfaces for Musical Expression*, 59–63.

Jordà, S. (2005). *Digital Lutherie Crafting Musical Computers for New Musics' Performance and Improvisation* [PhD thesis]. Universitat Pompeu Fabra, Barcelona.

Jourdan, T., and Caramiaux, B. (2023). "Culture and politics of machine learning in NIME: a preliminary qualitative inquiry," in *Proceedings of the International Conference on New Interfaces for Musical Expression*.

Kane, B. (2007). L'objet sonore maintenant: Pierre schaeffer, sound objects and the phenomenological reduction. *Organised Sound* 12, 15–24. doi: 10.1017/S135577180700163X

Keebler, J. R., Wiltshire, T. J., Smith, D. C., Fiore, S. M., and Bedwell, J. S. (2014). Shifting the paradigm of music instruction: implications of embodiment stemming from an augmented reality guitar learning system. *Front. Psychol.* 5:471. doi: 10.3389/fpsyg.2014.00471

Kirby, J. (2023). Approaches for working with the body in the design of electronic music performance systems. *Contemp. Music Rev.* 42, 304–318. doi: 10.1080/07494467.2023.2277564

Kirsh, D. (2013). Embodied cognition and the magical future of interaction design. *ACM Trans. Comput.-Hum. Interact.* 20, 1–30. doi: 10.1145/2442106.2442109

Knapp, R. B., and Lusted, H. S. (1990). A bioelectric controller for computer music applications. *Comput. Music J.* 14, 42–47. doi: 10.2307/3680115

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., et al. (2023). "Audiogen: textually guided audio generation," in *Proceedings of the 11th International Conference on Learning Representations*.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., et al. (2019). "MelGAN: generative adversarial networks for conditional waveform synthesis," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (Vancouver, BC), 12.

Küssner, M. B., Tidhar, D., Prior, H. M., and Leech-Wilkinson, D. (2014). Musicians are more consistent: gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Front. Psychol.* 5:789. doi: 10.3389/fpsyg.2014.00789

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, MA: Harvard UP.

Leman, M., and Maes, P.-J. (2014). The role of embodiment in the perception of music. *Empir. Musicol. Rev.* 9, 236–246. doi: 10.18061/emr.v9i3-4.4498

Leonard, J., Cadoz, C., Castagné, N., Florens, J.-L., and Luciani, A. (2014). "A virtual reality platform for musical creation: genesis-RT," in *Sound, Music, and Motion: 10th International Symposium, CMMR 2013, Marseille, France, October 15-18, 2013. Revised Selected Papers 10* (Cham: Springer), 346–371. doi: 10.1007/978-3-319-12976-1_22

Lindeman, R. W., Sibert, J. L., and Hahn, J. K. (1999). "Hand-held windows: towards effective 2D interaction in immersive virtual environments," in *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)* (Houston, TX: IEEE), 205–212. doi: 10.1109/VR.1999.756952

Ma, Y., Øland, A., Ragni, A., Sette, B. M. D., Saitis, C., Donahue, C., et al. (2024). Foundation models for music: a survey. *arXiv* [Preprint] arXiv:2408.14340. doi: 10.48550/arXiv.2408.14340

Magnusson, T. (2009). Of epistemic tools: musical instruments as cognitive extensions. *Organised Sound* 14, 168–176. doi: 10.1017/S1355771809000272

Magnusson, T. (2023). "Creative elenctics: playing with intelligent instruments," in *Invited Talk at the Computer Human Interaction and Music nEtwork (CHIME) Seminar Series*. Available online at: https://www.chime.ac.uk/media

Magnusson, T., Kiefer, C., and Ulfarsson, H. (2022). "Reflexions upon feedback," in *Proceedings of the International Conference on New Interfaces for Musical Expression*. PubPub. doi: 10.21428/92fbeb44.aa7de712

Mäki-Patola, T., Laitinen, J., Kanerva, A., and Takala, T. (2005). "Experiments with virtual reality instruments," in *Proceedings of the 2005 Conference on New Interfaces for Musical Expression*, 11–16.

Malloch, J., and Wanderley, M. M. (2017). "Embodied cognition and digital musical instruments: design and performance," in *The Routledge Companion to Embodied Music Interaction* (London: Routledge), 438–447. doi: 10.4324/9781315621364-48

Martinez Avila, J., Hazzard, A., Greenhalgh, C., Benford, S., and McPherson, A. (2023). The stretchy strap: supporting encumbered interaction with guitars. *J. New Music Res.* 52, 19–40. doi: 10.1080/09298215.2023.2274832

Martinez Avila, J. P., Tsaknaki, V., Karpashevich, P., Windlin, C., Valenti, N., Höök, K., et al. (2020). "Soma design for NIME," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 489–494.

Masu, R., Conci, A., Menestrina, Z., Morreale, F., and De Angeli, A. (2016). "Beatfield: an open-meaning audiovisual exploration," in *COOP 2016: Proceedings of the 12th International Conference on the Design of Cooperative Systems, 23-27 May 2016, Trento, Italy* (Cham: Springer). doi: 10.1007/978-3-319-33464-6_20

McDermott, J., Gifford, T., Bouwer, A., and Wagy, M. (2013). "Should music interaction be easy?" in *Music and Human-Computer Interaction* (Cham: Springer), 29–47. doi: 10.1007/978-1-4471-2990-5_2

McMillan, A., and Morreale, F. (2023). Designing accessible musical instruments by addressing musician-instrument relationships. *Front. Comput. Sci.* 5:1153232. doi: 10.3389/fcomp.2023.1153232

McPherson, A. (2017). Bela: an embedded platform for low-latency feedback control of sound. *J. Acoust. Soc. Am.* 141(5_Supplement):3618. doi: 10.1121/1.4987761

McPherson, A. P., Jack, R. H., and Moro, G. (2016). "Action-sound latency: are our tools fast enough?" in *Proceedings of the International Conference on New Interfaces for Musical Expression* (Brisbane City: Griffith University), 20–25.

McPherson, A. P., and Zappi, V. (2015). "Exposing the scaffolding of digital instruments with hardware-software feedback loops," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 162–167.

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., et al. (2016). Samplernn: an unconditional end-to-end neural audio generation model. *arXiv* [Preprint]. arXiv:1612.07837. doi: 10.48550/arXiv.1612.07837

Merleau-Ponty, M., Landes, D., Carman, T., and Lefort, C. (2013). *Phenomenology of Perception*. London: Routledge. doi: 10.4324/9780203720714

Mice, L., and McPherson, A. P. (2021). "Embodied cognition in performers of large acoustic instruments as a method of designing new large digital musical instruments," in *Proceedings of the 14th International Symposium on Computer Music Modeling and Retrieval* (Cham: Springer), 577–590. doi: 10.1007/978-3-030-70210-6_37

Moore, F. R. (1988). The dysfunctions of midi. *Comput. Music J.* 12, 19–28. doi: 10.2307/3679834

Moore, T. (2024). "Musical agents, agency, & AI: towards a phenomenological understanding," in *Proceedings of the International Computer Music Conference*, 102–105.

Moral-Bofill, L., López de la Llave, A., Pérez-Llantada, M. C., and Holgado-Tello, F. P. (2022). Development of flow state self-regulation skills and coping with musical performance anxiety: design and evaluation of an electronically implemented psychological program. *Front. Psychol.* 13:899621. doi: 10.3389/fpsyg.2022.899621

Morreale, F., McPherson, A., and Wanderley, M. (2018). "NIME identity from the performer?s perspective," in *Proceedings of the International Conference on New Interfaces for Musical Expression*.

Morrison, L., and McPherson, A. (2024). "Entangling entanglement: a diffractive dialogue on hci and musical interactions," in *Proceedings of the CHI*

*Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–17. doi: 10.1145/3613904.3642171

Nash, C. (2011). *Supporting Virtuosity and Flow in Computer Music* [PhD thesis]. Sanit John's College, University of Cambridge, Cambridge.

Nercessian, S., McClellan, R., Goldsmith, C., Fink, A. M., and LaPenn, N. (2023). "Real-time singing voice conversion plug-in," in *Proceedings of the 26th International Conference on Digital Audio Effects*, 351–354.

Neupert, M., and Wegener, C. (2019). "Isochronous control+ audio streams for acoustic interfaces," in *Proceedings of the 17th Linux Audio Conference*, 5.

Nick, C. (2007). "Live coding practice," in *Proceedings of the 7th International Conference on New Interfaces for Musical Expression* (New York, NY: ACM), 112–117. doi: 10.1145/1279740.1279760

Nierhaus, G. (2009). *Algorithmic Composition: Paradigms of Automated Music Generation*. Cham: Springer Science & Business Media.

Odom, W. (2024). "Illustrating, annotating & extending design qualities of slow technology," in *Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction, NordiCHI '24 Adjunct* (New York, NY: Association for Computing Machinery), 1. doi: 10.1145/3677045.3685499

O'Modhrain, M. S. (2001). *Playing by Feel: Incorporating Haptic Feedback into Computer-based Musical Instruments* [PhD thesis]. Stanford University, Stanford, CA.

O'Modhrain, S. (2011). A framework for the evaluation of digital musical instruments. *Comput. Music J.* 35, 28–42. doi: 10.1162/COMJ_a_00038

O'Modhrain, S., and Gillespie, R. B. (2018). "Once more, with feeling: revisiting the role of touch in performer-instrument interaction," in *Musical Haptics*, eds. S. Papetti, and C. Saitis (Cham: Springer), 11–27. doi: 10.1007/978-3-319-58316-7_2

Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016a). Wavenet: a generative model for raw audio. *arXiv [Preprint]* arXiv:1609.03499. doi: 10.48550/arXiv.1609.03499

Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., Kavukcuoglu, K. (2016b). Conditional image generation with PixelCNN decoders. *arXiv* [Preprint] arXiv:1606.05328. doi: 10.48550/arXiv.1606.05328

Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., et al. (2017). Parallel WaveNet: fast high-fidelity speech synthesis. *arXiv* [Preprint] arXiv:1711.10433. doi: 10.48550/arXiv.1711.10433

Oore, S. (2005). "Learning advanced skills on new instruments," in *Proceedings of the International Conference on New interfaces for Musical Expression*, 60–64.

Pelinski, T., Shepardson, V., Symons, S., Caspe, F. S., Temprano, A. L. B., Armitage, J., et al. (2022). "Embedded AI for NIME: challenges and opportunities," in *Proceedings of the International Conference on New Interfaces for Musical Expression*. PubPub. doi: 10.21428/92fbeb44.76beab02

Privato, N., Lepri, G., Magnusson, T., and Einarsson, E. T. (2024). "Sketiching magnetic interactions for neural synthesis," in *Proceedings of the International Conference on Technologies for Music Notation and Representation*, 89–97.

Privato, N., Magnusson, T., and Einarsson, E. T. (2023). "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*, 36–44.

Reed, C. N., Nordmoen, C., Martelloni, A., Lepri, G., Robson, N., Zayas-Garin, E., et al. (2022). "Exploring experiences with new musical instruments through micro-phenomenology," in *Proceedings of the International Conference on New Interfaces for Musical Expression*. doi: 10.21428/92fbeb44.b304e4b1

Roberts, C., and Wakefield, G. (2018). "Tensions and techniques in live coding performance," in *The Oxford Handbook of Algorithmic Music* (Oxford: Oxford University Press), 293–318. doi: 10.1093/oxfordhb/9780190226992.013.20

Russell, S. J., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Saddle River, NJ: Prentice Hall.

Russolo, L. (1913). *The Art of Noise. A Great Bear Pamphlet*. New York, NY: Something Else Press.

Ryan, J. (1991). Some remarks on musical instrument design at steim. *Contemp. Music Rev.* 6, 3–17. doi: 10.1080/07494469100640021

Saitis, C., Järveläinen, H., and Fritz, C. (2018). "The role of haptic cues in musical instrument quality perception," in *Musical Haptics*, eds. S. Papetti, and C. Saitis (Cham: Springer International Publishing), 73–93. doi: 10.1007/978-3-319-58316-7_5

Sayer, T. (2016). Cognitive load and live coding: a comparison with improvisation using traditional instruments. *Int. J. Performance Arts Digit. Media* 12, 129–138. doi: 10.1080/14794713.2016.1227603

Schaeffer, P. (1964). *Traité des Objets Musicaux*. Seuil, nouv. éd edition edition. Paris: Éditions du Seuil.

Schafer, R. M. (1977). *The Tuning of the World*. New York, NY: Random House Inc.

Serafin, S., Erkut, C., Kojs, J., Nilsson, N. C., and Nordahl, R. (2016). Virtual reality musical instruments: State of the art, design principles, and future directions. *Comput. Music J.* 40, 22–40. doi: 10.1162/COMJ_a_00372

Shepardson, V., Reus, J., and Magnusson, T. (2024). "Tungnaá: a hyper-realistic voice synthesis instrument for real-time exploration of extended vocal expressions," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 536–540.

Shier, J., Saitis, C., Robertson, A., and McPherson, A. (2024). "Real-time timbre remapping with differentiable DSP," in *Proceedings of the International Conference on New Interfaces for Musical Expression*.

Shusterman, R. (2008). *Body Consciousness: A Philosophy of Mindfulness and Somaesthetics*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511802829

Sivanandam, S. N., and Deepa, S. N. (2007). *Introduction to Genetic Algorithms*. Berlin: Springer.

Smalley, D. (1997). Spectromorphology: explaining sound-shapes. *Organised Sound* 2, 107–126. doi: 10.1017/S1355771897009059

Sonami, L. (2024). "REPETITION and DESIRE: echo, narcissus, AI and I," in *Keynote at the International Computer Music Conference 2024* (Seoul). Available online at: https://sonami.net/wp-content/uploads/2024/07/SOnami-ICMC24-KEYnOTE.pdf

Stockhausen, K. (1972). *Four Criteria of Electronic Music with Examples from Kontakte*. London: Marion Boyars Publishers.

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, MA: Cambridge university press.

Tanaka, A. (2010). "Mapping out instruments, affordances, and mobiles," in *Proceedings of the International Conference on New interfaces for Musical Expression*, 88–93.

Tapparo, C. S., and Zappi, V. (2022). "Bodily awareness through NIMEs: deautomatising music making processes," in *Proceedings of the International Conference on New Interfaces for Musical Expression*. PubPub. doi: 10.21428/92fbeb44.7e04cfc8

Tatar, K. (2019). *Musical Agents Based on Self-organizing Maps for Audio Applications* [Thesis]. Communication, Art & Technology: School of Interactive Arts and Technology, Surrey, BC.

Tatar, K., Bisig, D., and Pasquier, P. (2020). Latent timbre synthesis. *Neural Comput. Appl.* 33, 67–84. doi: 10.1007/s00521-020-05424-2

Tatar, K., Cotton, K., and Bisig, D. (2023). Sound design strategies for latent audio space explorations using deep learning architectures. *arXiv* [Preprint] arXiv:2305.15571. doi: 10.48550/arXiv.2305.15571

Tatar, K., and Pasquier, P. (2019). Musical agents: a typology and state of the art towards Musical Metacreation. *J. New Music Res.* 48, 56–105. doi: 10.1080/09298215.2018.1511736

Tomczak, J. M. (2022). *Deep Generative Modeling*. Cham: Springer International Publishing. doi: 10.1007/978-3-030-93158-2

Tragtenberg, J., Calegario, F., Wanderley, M. M., and Cavalcanti, V. (2024). "Designing DMIS with (in) a music culture: a participatory design process with the xambá quilombola community," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 367–376.

Turian, J., Shier, J., Tzanetakis, G., McNally, K., and Henry, M. (2021). one billion audio sounds from GPU-enabled modular synthesis. *arXiv* [Preprint]. arXiv:2104.12922. doi: 10.48550/arXiv.2104.12922

Tuuri, K., and Eerola, T. (2012). Formulating a revised taxonomy for modes of listening. *J. New Music Res.* 41, 137–152. doi: 10.1080/09298215.2011.614951

Varela, F. J., Thompson, E., and Rosch, E. (2017). *The Embodied Mind, Revised Edition: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262529365.001.0001

Varèse, E., and Wen-chung, C. (1966). The liberation of sound. *Perspect. New Music* 5, 11–19. doi: 10.2307/832385

Wessel, D., and Wright, M. (2002). Problems and prospects for intimate musical control of computers. *Comput. Music J.* 26, 11–22. doi: 10.1162/014892602320582945

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*. Hoboken, NJ: John Wiley & Sons.

Wright, A., Damskägg, E.-P., Juvela, L., and Välimäki, V. (2020). Real-time guitar amplifier emulation with deep learning. *Appl. Sci.* 10:766. doi: 10.3390/app10030766

Wrigley, W. J., and Emmerson, S. B. (2013). The experience of the flow state in live music performance. *Psychol. Music* 41, 292–305. doi: 10.1177/0305735611425903

Yang, L.-C., Chou, S.-Y., and Yang, Y.-H. (2019). "Midinet: a convolutional generative adversarial network for symbolic-domain music generation," in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 324–331.

Young, G. W., Murphy, D., and Weeter, J. (2018). "A functional analysis of haptic feedback in digital musical instrument interactions," in *Musical Haptics*,

eds. S. Papetti, and C. Saitis (Cham: Springer), 95–122. doi: 10.1007/978-3-319-583 16-7_6

Zamborlin, B. (2015). *Studies on Customisation-driven Digital Music Instruments* [PhD thesis]. Goldsmiths, University of London, London.

Zappi, V., and McPherson, A. (2018). Hackable instruments: supporting appropriation and modification in digital musical interaction. *Front. ICT* 5:26. doi: 10.3389/fict.2018.00026

Zappi, V., and McPherson, A. P. (2014). "Dimensionality and appropriation in digital musical instrument design," in *Proceedings of the International Conference on New Interfaces for Musical Expression, Vol. 14* (Utrecht: Citeseer), 455–460.

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *Proceedings of the 13th European Conference on Computer Vision, Part I* (Cham: Springer), 818–833. doi: 10.1007/978-3-319-10590-1_53