From Tacit to Tokenised: Critical Perspectives on AI Text-to-Audio Interfaces

CHARALAMPOS SAITIS, Queen Mary University of London, United Kingdom
FABIO MORREALE, Sony AI, Spain
LANDON MORRISON, University of Rochester, United States
ASHLEY NOEL-HIRST, Queen Mary University of London, United Kingdom
REBECCA FIEBRINK, University of the Arts London, United Kingdom
JOSEPH MEYER, University of the Arts London, United Kingdom
COURTNEY N. REED, Loughborough University London, United Kingdom

Abstract / Workshop Description

In AI text-to-audio interfaces, forms of tacit musical knowledge traditionally developed through embodied interaction are translated into tokenised linguistic descriptors that condition generative models. These models assume that language can serve as a proxy for structuring latent representations and producing musically meaningful outputs. However, musicians and researchers have long argued that music resists containment through verbal description, that it does not operate like a language and exceeds the limits of pure structure. This half-day workshop examines the functionalising of language in semantic cues for AI audio engines, which is moreover largely curated by a small number of multi-national industry giants, and its implications for digital musical instrument design and musical practice more broadly. Participants will engage critically with popular AI text-to-audio models alongside movement-to-audio interfaces to explore how alternative semantic mediation strategies for generating sound shape musical meaning and creativity. A collective open discussion will critically reflect on their cultural, epistemological, musical, aesthetic, and political-economic implications for the NIME and off-NIME communities.

Additional Key Words and Phrases: Critical AI, Text-to-Audio Generation, Movement-to-Audio Generation

1 Motivation

Recent advances in generative AI (GenAI) have accelerated the integration of natural language interfaces into artificial musical creative systems, allowing users to generate and manipulate sound through textual description. Text-to-audio generative models like Stable Audio Open (open source [6]) and Suno v5 (closed source) operate on the assumption that language provides a suitable proxy and evaluative mechanism for developing latent representations and generating musically/sonically meaningful outputs. Beyond GenAI, many music information retrieval tasks work on this assumption, from semantic audio feature extraction to automatic genre tagging [11, 31].

Conversely, musicians and researchers have long pointed to the limitations of language in both shaping and articulating creative practice [17, 19, 21]. This highlights a friction between tokenised linguistic interaction with generative audio tools and the ineffable and tacit, embodied forms of knowledge involved in music-making [1, 10]. This workshop is therefore motivated by the need to critically examine the functionalising of language in semantic cues for AI audio generation/synthesis, and the inevitable incorporation of such technology into Digital Musical Instrument (DMI) design. In doing so, the workshop aims to consolidate ongoing critical discussions on AI practices within the NIME and related communities [9, 14, 15, 18, 23]. Specifically, this workshop asks:

How does the functionalization of language in AI differ from the linguistic mediation of music in other contexts? For instance, analogue synths still have names attached to knobs, but these terms typically aim to describe some measurable (and thus controllable) parameter of sound (e.g., cutoff frequency for filtering). By comparison, the language of AI text-to-sound applications typically resides at a more vernacular level of human culture, so users are encouraged to prompt with the names of instruments, genres, and so on; turning these terms into control knobs clearly involves a more drastic level of conceptual reduction, flattening otherwise thick social/physical entities into a fixed, one-dimensional space. Several of the

Authors' Contact Information: Charalampos Saitis, Queen Mary University of London, City, United Kingdom; Fabio Morreale, Sony AI, City, Spain; Landon Morrison, University of Rochester, City, United States; Ashley Noel-Hirst, Queen Mary University of London, City, United Kingdom; Rebecca Fiebrink, University of the Arts London, City, United Kingdom; Joseph Meyer, University of the Arts London, City, United Kingdom; Courtney N. Reed, Loughborough University London, City, United Kingdom.

NIME '26, June 23–26, 2026, London, UK

workshop organisers have previously explored similar critiques of conceptual compression/reduction of culturally nuanced and ambiguous musical concepts by the data-driven representations that enable generative AI music/audio systems, such as timbre [27, 28], genre [3], and instrument [25]. Reflecting on these criticalities is compounded by a further observation, relevant to this year's conference theme:

While AI text-to-audio applications attempt to fix relations between words and sounds, there remains vast variety in the way language is applied to music across history and in different sonic cultures [7, 26]. So, what are the stakes of imposing a fixed relation onto the cultural/semantic tendencies of different communities of practice? How well do AI text-to-audio systems, and/or the data that train them, accommodate local/contextual meanings for users in different sonic cultures [22, 30]? To what extent will the text-sound mappings propagated by AI begin to bend the practices of these communities to its own representations? And how is this dynamic inflected by the extreme political-economic power that multi-national companies have over local communities of practice—e.g., through the mass production and dissemination of their tools, including assembling and curating massive datasets [20]?

The proposed workshop will probe these questions by combining hands-on exploration of input modalities in GenAI systems with critical discussion about input-to-sound mappings and construction of generated audio. Specifically, participants will engage with two alternative semantic mediation strategies for generative sound: natural language and embodied gesture. Contrasting text-to-sound and gesture-to-sound generation is particularly meaningful for NIME, where gesture has long been central to questions of embodiment and mapping [12, 29], and can illuminate broader epistemological tensions about how musical meaning is mediated and enacted. We foreground first- and second-person methodologies, inviting researchers to examine musical interaction design and AI practices from within, positioning their own experiences and creative processes as sites of inquiry [15], which can generate critical insight into how these technologies are encountered and negotiated in NIME and broader HCI contexts.

2 Workshop Structure

The workshop will span three hours and activities before, during, and following the workshop will be organised as detailed in the following sections (see also Table 1):

2.1 Pre Workshop Activities

Several preparatory activities will be undertaken prior to the workshop.

First, the organisers will conduct internally a pilot "mini-workshop" using the proposed tools and structure, which will function as a form of collaborative autoethnographic inquiry [2, 5]. Insights, tensions, or unexpected outcomes from this activity will inform refinements to the workshop design and discussion prompts.

Second, prospective participants will be invited to submit a brief expression of interest (150–200 words) no later than two weeks prior to the workshop, outlining their background and motivations for attending (see Section 5).

Third, confirmed participants will receive instructions for remote technical preparation, including a guided tutorial of of the text-to-sound and gesture-to-sound systems that will be used during the workshop. This may involve creating necessary accounts, downloading required software, and verifying compatibility in advance (e.g., setting up a HuggingFace account to access the Stable Audio Open encoder

The workshop will also endeavour to collect perspectives of interaction with AI systems (see Section 2.3). Therefore, we will share the full workshop itinerary and obtain informed consent (see Section 7) from participants to support ethnographic study during the workshop, including collecting observational accounts, taking photographs, and recording performances with designed audio.

2.2 Workshop Activities

The workshop will begin with an introductory session led by the organisers, who will briefly present their backgrounds and outline the conceptual motivations and objectives of the workshop. We will also review the schedule and structure for the workshop at this time. An ice-breaking activity will follow, inviting participants to introduce themselves to others, in a fun and memorable way. After the icebreaker, we will review ethics considerations and consent provided in the pre-workshop stage to ensure continued inform consent for data collection. We will then assist participants in finalising their technical setups.

Participants will form pairs and be divided into two groups, each working with one input modality for 40 min before switching for another 40 min, ensuring that all participants explore both semantic mediation strategies individually. Working in pairs supports both first- and second-person inquiry by allowing participants to reflect on their own creative processes while simultaneously engaging with the perspectives and interpretations of another practitioner. It also foregrounds communication and negotiation between different sonic vocabularies and communities of practice, which aligns with this year's NIME theme.

Table 1. Planned workshop schedule, including pre and post workshop activities

Activities		Timeline/Duration
Pre Workshop	Pilot mini-workshop + collaborative autoethnography	by end May
	prospective participants apply	by 9 June
	remote technical preparation	19 + 22 June
Workshop	Introduction + ice breaker + consent	20 min
	Technical setup check + guided tutorial (optional)	10 min
	Input-to-sound exploration I	40 min
	Input-to-sound exploration II (swap input modality)	40 min
	Sounds/performances + reflective discussion	50 min
	Closing remarks + networking	10 min
Post Workshop	Curated online outlet shared with participants and community	by end July

In terms of natural language, participants will interact with a lightweight text-conditioned generative audio system (e.g., RAVE [4] or Stable Audio Open), using any text descriptions (prompts) that feel natural to them / any descriptions they choose / including narratives or words we suggest, processed by pre-trained text embeddings (e.g., Stable Audio Open's or CLAP [32]). In terms of embodied gesture, participants will use movements that feel natural to them / any movements they choose / including choreographies or poses we suggest, detected by their laptop camera and processed by pre-trained pose estimators (e.g., PoseNet [16] or MoveNet [13]) connected to a lightweight gesture-to-audio generative system (also using RAVE or Stable Audio Open).

Using browser-run or Max/Pd interfaces including ones developed as part of Noel-Hirst and Meyer's own PhD researches, and ones built on Fiebrink's Wekinator [8], participants will train each system in a supervised way by associating textual/gestural inputs with some audio (e.g. vocalisations, found sounds, own samples, or open license samples, including options we provide). They may also train unsupervised input-to-sound generative systems by inputing some text / demonstrating some movements and uploading some audio for the system to learn a mapping between.

The emphasis will be on critical artistic exploration through rapid prototyping rather than refinement. We will encourage idiosyncratic sonic vocabularies rather than polished outputs, guided by briefs such as "create some sounds for a performance you are about to give tonight" or "make the most foolish sound you can imagine" or "craft a sound that might be physically implausible." Participants will be guided to explore text descriptions that progress from literal to more fictional and gestures that vary from conventional to more abstract. They may also explore how input-to-sound relations propagated by AI translate between their respective practices and between input modalities (e.g., attempting to recreate each other's text- or movement-generated sounds via own prompting).

We will facilitate a shared ideation canvas (e.g., using Miro[1] or FigJam[2]) in which participants will document their explorations, including strategies, semantic compressions or distortions, breakdowns or unexpected alignments, encouraging reflections on the questions posed by the workshop. Following the exploratory activities, the full group will have a collective open discussion on the workshop's questions, with performances of designed audio and the results of the canvas as conversation starters. Organisers will guide reflective inquiry through a diverse set of disciplinary lenses (see Section 3), introducing deliberate provocations where appropriate [24]. We will conclude the workshop with a networking activity, in which participants will have a chance to break into conversation and share contact details.

2.3 Post Workshop Activities

Following the workshop, we will develop and circulate a curated online outlet (e.g., a dedicated page on the workshop's website or blog post) documenting materials and insights generated during the workshop. This will include selected artefacts such as audio excerpts, short video clips, photographs or screenshots, sketches, and code snippets, alongside a synthesis of findings derived from a reflexive thematic analysis of participants' explorations and discussions.

The aim of this documentation is not merely archival, but reflective: to surface recurring themes, tensions, and design implications that emerged across the workshop activities. The curated output will be shared with workshop participants for feedback prior to public release and subsequently disseminated to the broader NIME community, contributing to ongoing conversations around AI practices and research within DMI design and HCI more broadly.

[1]https://miro.com/

[2]https://www.figma.com/figjam/

3 Organisers

The workshop will be organized by researchers who work at the intersection of Music, AI, HCI, and NIME. They bring together perspectives from timbre research, human–computer interaction, instrument design, creative machine learning and AI, performance and artistic practice, and critical and cultural approaches to music and technology.

Charalampos Saitis (he/him) is an Assistant Professor of Digital Music Processing at the Centre for Digital Music at Queen Mary University of London. His work combines methods from psychoacoustics, music informatics, and human–computer interaction to inform the design of new musical interfaces and AI-driven creative tools, with a particular focus on how timbre is conceptualised in the design and use of new music technologies and instruments. He has published widely on timbre perception and semantics and has been actively involved in the NIME and ISMIR communities.

Fabio Morreale (he/him) is a scholar and musician working in the intersection of NIME, AI, Ethics, and Philosophy of Technology. He has an MS in Computer Science, a PhD in HCI, and now works as a Staff Research Scientist at Sony AI. Prior to that, he was a Senior Lecturer in Algorithmic Composition at the University of Auckland in New Zealand. He also worked as a postdoc at Queen Mary University of London in the Augmented Instruments Lab. He has been publishing at NIME since 2014 and is an active member of the community.

Landon Morrison (he/him) is an Assistant Professor of Music Theory at the Eastman School of Music at University of Rochester, New York. He has previously worked as a post-doctoral researcher at Imperial College London and as a college fellow and lecturer at Harvard University. His research, which has been published in leading academic journals, draws on methods from music and media studies to examine technocultural mediation in contemporary sonic practices.

Ashley Noel-Hirst (he/him) is an artist and fourth-year PhD student at the Centre for Digital Music at Queen Mary University of London. He has an MMus in Sonic Arts from Goldsmiths University of London and has created interactive and generative music for several festivals, including SXSW (US) and Being Human (UK). In his research, Ashley draws on AI, HCI and ethnomusicology to investigate the impact of timbre representations in AI systems on sample-based music-making.

Rebecca Fiebrink (she/her) is Professor of Creative Computing at the Creative Computing Institute at University of the Arts London. She has been conducting research on integrating machine learning and AI in human creative practices, including the design of new musical instruments, for over twenty years. She is the creator of numerous tools for creative and embodied machine learning, including Wekinator [8], which have tens of thousands of users. She has organised popular workshops at past NIME conferences as well as at ISMIR, CHI, IUI, NeurIPS, and others.

Joseph Meyer (they/their) is a second-year PhD student and Associate Lecturer at UAL's Creative Computing Institute, researching interactive neural audio synthesis and movement-to-sound mapping. Their work emphasizes techniques to lower barriers to entry for artists to access and use generative ML systems. Recently, they have been developing methods to make generative models more controllable and interactive, and to increase the novelty of generated output.

Courtney N. Reed (she/her) is an Assistant Professor of Digital Technologies and Creative Futures at Loughborough University London. Her work focuses on vocalist-voice relationships and how singers conceptualise their bodies and voice in the design and use of digital musical instruments. She is an expert in subjective methodologies and experiential querying and has previously led workshops on sensory experiences at NIME, TEI, and CHI. She is especially fascinated by ambiguity and metaphors in working with and communicating abstract, embodied sensory experiences.

4 Technical and Space Requirements

We expect to host on-site about 20 participants, aiming for a diversity of communities, practices, and career stages, fostering a rich and plural dialogue aligned with the workshop's aims. We require a room with seating capacity for about 27 individuals in total, including the organisers, arranged in a format conducive to small-group collaboration and hands-on experimentation. The preferred format for the workshop is fully in-person; however, a small number of participants joining remotely can be accommodated.

Due to the nature of the workshop, the space must provide a reliable internet connection to support real-time interaction with cloud-based AI tools, as well as essential audio equipment including a PA system with speakers, a digital audio interface, a small mixer, and at least one microphone. On-site support for AV equipment configuration would be welcomed; however, the organisers, including one who works at the host institution, are prepared to manage setup and troubleshooting if necessary.

All workshop-specific materials, including access to AI models and shared digital collaboration platforms (e.g., Miro boards or equivalent), will be provided by the organising team. Participants will be informed in advance of any personal equipment they are encouraged to bring (e.g., laptops, headphones, own sounds), but no specialised hardware will be required beyond what is supplied.

5 Call for Participation

We invite musicians, researchers, artists, designers, educators, and creative practitioners from NIME and off-NIME communities, who are interested in representations of sound and how AI systems learn and interpret text-based descriptions of audio. This half-day workshop will explore, through hands-on activities and collective discussion, how text-to-sound systems attempt to fix relations between words and sound, and how these mappings intersect with the diverse sonic cultures, languages, pedagogies, and histories that shape musical practice. In doing so, we will reflect on how we talk about music, how we negotiate audio representations, and the values we can construct in AI systems, considering the global implications and hegemonies being introduced by multi-national AI giants in shaping the way we conceptualise sound.

Interested attendees should submit a brief expression of interest (150–200 words) no later than two weeks prior to the workshop, outlining their background and motivation for attending. Our aim is to curate a diverse group of participants representing a range of perspectives and practices. As part of the workshop, the organisers will conduct an ethnographic study of participants' experiences and reflections, including observational notes, contributions to a shared ideation canvas, photographs of the workshop process, and recordings of short performances with generated audio. Participation in data collection will be voluntary and subject to informed consent.

6 Media Links

We will use a dedicated website[3] to advertise the workshop to a wide audience, to publish the workshop description and proposal, planned activities, and schedule. Our Call for Participation will point towards this website.

7 Ethical Standards and Open Research Practices

The workshop will adhere to the standards and guidelines established by NIME's ethics, environmental, and diversity committees. Upon acceptance of the proposal, formal ethical approval for the planned activities will be sought from the lead organiser's institution (Queen Mary University of London) and, where required, from the institutions of co-organisers. All workshop procedures that involve data collection and documentation will be conducted in accordance with institutional research ethics frameworks and relevant data protection regulations.

Wherever possible, we will ensure research data used and generated in the workshop, including digital and rich media outputs, is made openly available under Creative Commons CC-BY licenses, except where this would be inappropriate due to e.g. reasons of personal data protection, participant consent, IP/commercial confidentiality or inclusion of other copyright material, for example, sensitive artistic processes or personal narratives.

References

[1] Carolyn Abbate. 2004. Music—Drastic or Gnostic? *Critical Inquiry* 30, 3 (March 2004), 505–536. https://doi.org/10.1086/421160

[2] Paulo Bala, Pedro Sanches, Vanessa Cesário, Sarah Leão, Catarina Rodrigues, Nuno Jardim Nunes, and Valentina Nisi. 2023. Towards Critical Heritage in the wild: Analysing Discomfort through Collaborative Autoethnography. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 771, 19 pages. https://doi.org/10.1145/3544548.3581274

[3] Nick Bryan-Kinns, Ashley Noel-Hirst, and Corey Ford. 2024. Using Incongruous Genres to Explore Music Making with AI Generated Content. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C&C '24)*. Association for Computing Machinery, New York, NY, USA, 229–240. https://doi.org/10.1145/3635636.3656198

[4] Antoine Caillon and Philippe Esling. 2022. Streamable Neural Audio Synthesis With Non-Causal Convolutions. In *International Conference on Digital Audio Effects (DaFX)*. 1–7.

[5] Carolyn Ellis, Tony E Adams, and Arthur P Bochner. 2011. Autoethnography: An Overview. *Historical Social Research / Historische Sozialforschung* 4 (2011), 273–290.

[6] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable Audio Open. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[7] Steven Feld and Aaron A Fox. 1994. Music and Language. *Annual Review of Anthropology* (1994), 25–53.

[8] Rebecca Fiebrink and Perry R Cook. 2010. The Wekinator: A system for real-time, interactive machine learning in music. In *Proceedings of the Eleventh International Society for Music Information Retrieval Conference (ISMIR)*.

[9] Rebecca Fiebrink and Laetitia Sonami. 2020. Reflections on Eight Years of Instrument Creation with Machine Learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 237–242. https://doi.org/10.5281/zenodo.4813334

[10] Michael Gallope. 2019. *Deep Refrains: Music, Philosophy, and the Ineffable*. University of Chicago Press.

[11] Owen Green, Bob LT Sturm, Georgina Born, and Melanie Wald-Fuhrmann. 2024. A critical survey of research in music genre recognition. In *Proceedngs of the 25th International Society for Music Information Retrieval Conference (ISMIR)*. 1–38.

[12] Alexander Refsum Jensenius. 2022. *Sound Actions: Conceptualizing Musical Instruments*. MIT Press.

[13] Rishabh Bajpaiand Deepak Joshi and Deepak Joshi. 2021. MoveNet: A Deep Neural Network for Joint Profile Prediction Across Variable Walking Speeds and Slopes. *IEEE Transactions on Instrumentation and Measurement* 70 (2021), 1–11. https://doi.org/10.1109/tim.2021.3073720

[3] https://comma.eecs.qmul.ac.uk/critical-text-audio-interfaces/

[14] Théo Jourdan and Baptiste Caramiaux. 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*, Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Mexico City, Mexico, Article 47, 7 pages. https://doi.org/10.5281/zenodo.11189202

[15] Théo Jourdan, T Pelinski Ramos, Hugo Scurto, Kristina Höök, et al. 2024. First-person and second-person perspectives for ML in NIME. In *Workshop at the International Conference on New Interfaces for Musical Expression*.

[16] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, USA, 2938–2946. https://doi.org/10.1109/ICCV.2015.336

[17] David Lidov. 2005. *Is language a music?: Writings on musical form and signification*. Indiana University Press.

[18] C Martin, Fabio Morreale, Benedikte Wallace, and Hugo Scurto. 2020. Critical perspectives on AI/ML in musical interfaces. In *Workshop at the International Conference on New Interfaces for Musical Expression*.

[19] Andrew McPherson, Landon Morrison, Matthew Davison, and Marcelo M. Wanderley. 2024. On mapping as a technoscientific practice in digital musical instruments. *Journal of New Music Research* 53, 1–2 (March 2024), 110–125. https://doi.org/10.1080/09298215.2024.2442356

[20] Fabio Morreale. 2021. Where does the buck stop? Ethical and political issues with AI in music creation. *Transactions of the International Society for Music Information Retrieval* 4, 1 (2021).

[21] Fabio Morreale, Marco A Martinez-Ramirez, Raul Masu, WeiHsiang Liao, and Yuki Mitsufuji. 2025. Reductive, exclusionary, normalising: the limits of generative AI music. *Transactions of the International Society for Music Information Retrieval* 8, 1 (2025).

[22] Fabio Morreale, Megha Sharma, I-Chieh Wei, et al. 2023. Data Collection in Music Generation Training Sets: A Critical Analysis.. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. 37–46.

[23] Ashley Noel-Hirst, Charalampos Saitis, and Nick Bryan-Kinns. 2025. Sampling the Latent Space: Exploring the Creative Potential of Generative AI Through the Lens of Sample-Based Music Making. In *Proceedings of The Sixth Conference on AI Music Creativity (AIMC)*.

[24] Dimitrios Raptis, Rikke Hagensby Jensen, Jesper Kjeldskov, and Mikael B. Skov. 2017. Aesthetic, Functional and Conceptual Provocation in Research Through Design. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. Association for Computing Machinery, New York, NY, USA, 29–41. https://doi.org/10.1145/3064663.3064739

[25] Courtney N. Reed, Adan L. Benito, Franco Caspe, and Andrew P. McPherson. 2024. Shifting Ambiguity, Collapsing Indeterminacy: Designing with Data as Baradian Apparatus. *ACM Trans. Comput.-Hum. Interact.* 31, 6, Article 73 (Dec. 2024), 41 pages. https://doi.org/10.1145/3689043

[26] Courtney N. Reed, Paul Strohmeier, and Andrew P. McPherson. 2023. Negotiating Experience and Communicating Information Through Abstract Metaphor. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 185, 16 pages. https://doi.org/10.1145/3544548.3580700

[27] Charalampos Saitis, Bleiz M Del Sette, Jordan Shier, Haokun Tian, Shuoyang Zheng, Sophie Skach, Courtney N. Reed, and Corey Ford. 2024. Timbre Tools: Ethnographic Perspectives on Timbre and Sonic Cultures in Hackathon Designs. In *Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures* (Milan, Italy) *(AM '24)*. Association for Computing Machinery, New York, NY, USA, 229–244. https://doi.org/10.1145/3678299.3678322

[28] Charalampos Saitis, Courtney N. Reed, Ashley Laurent Noel-Hirst, Giacomo Lepri, and Andrew McPherson. 2025. (De)Constructing Timbre at NIME: Reflecting on Technology and Aesthetic Entanglements in Instrument Design. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, 197–206. https://doi.org/10.5281/zenodo.15698835

[29] Koray Tahiroglu, Michael Gurevich, and R. Benjamin Knapp. 2018. Contextualising Idiomatic Gestures in Musical Interactions with NIMEs. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Thomas Martin Luke Dahl, Douglas Bowman (Ed.). Virginia Tech, Blacksburg, Virginia, USA, 126–131. https://doi.org/10.5281/zenodo.1302701

[30] Tina Tallon. 2024. Words About Music: the Promise and Peril of Text-to-Music Models. In *Proceedings of the CHIME Conference 2024*.

[31] Geraint A. Wiggins. 2009. Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. In *Proceedings of the 2009 11th IEEE International Symposium on Multimedia (ISM '09)*. IEEE Computer Society, USA, 477–482. https://doi.org/10.1109/ISM.2009.36

[32] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.