# Interactive Movement-to-Audio with Pre-Trained Neural Networks

Mapping Embodied Practice to Sound

JOSEPH MEYER
Creative Computing Institute, University of the Arts London, j.meyer@arts.ac.uk

Nick Bryan-Kinns
Creative Computing Institute, University of the Arts London, n.bryankinns@arts.ac.uk

Sarah Fdili Alaoui
Creative Computing Institute, University of the Arts London, s.fdilialaoui@arts.ac.uk

Mick Grierson
Creative Computing Institute, University of the Arts London, m.grierson@arts.ac.uk

Rebecca Fiebrink
Creative Computing Institute, University of the Arts London, r.fiebrink@arts.ac.uk

Systems to interactively generate audio from human movement are used by artists including dancers to support their performances and practice. However, current real-time movement-to-sound systems require specialized hardware or expertise, or map only very simple movement-to-audio relationships. We present a new technique and system implementation for interactive sonification of human movement through unsupervised machine learning. Our system maps between latent spaces, linking a pose estimator to a neural audio generator to enable sonification of human bodies. This may lower barriers to entry for artists to generate sound from their embodied movement through complex mappings. Our system requires no specialized hardware or niche AI expertise, minimal data to learn a user's custom movements, and trains extremely fast. It represents a new method for mapping custom data to a latent space through unsupervised learning, and advances state-of-the-art interactive movement sonification through its increased accessibility and ease of use relative to its complexity.

**CCS CONCEPTS** • Human-centered computing~Human computer interaction (HCI) • Computing methodologies~Machine learning • Applied computing~Arts and humanities~Sound and music computing • Computing methodologies~Machine learning~Learning paradigms~Unsupervised learning • Computing methodologies~Machine learning~Machine learning approaches~Neural networks

**ACM Reference Format:**

Joseph Meyer, Nick Bryan-Kinns, Sarah Fdili Alaoui, Mick Grierson, and Rebecca Fiebrink. 2025. Interactive Movement-to-Audio with Pre-Trained Neural Networks. In Creativity and Cognition (C&C '25), June 23–25, 2025, Virtual, United Kingdom. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3698061.3734415

# 1 Introduction

Movement-to-sound systems, a subset of creativity support tools [16], are used by dancers to dynamically generate soundtracks for their choreography or facilitate real-time sonification of their improvisations [3] [9] [13]. They are also used by music artists to produce novel audio samples from gestural input [12]. Additionally, academics are interested in movement-sound systems to examine relationships between modalities [2] [18].

However current architectures are either too simple to yield complex (i.e. implicit) mappings [10], too complicated to use without niche AI expertise [13], or too heavy to customize to individual users or interact with in real-time [17]. Directly mapping body part locations to synthesizer parameters facilitates real-time movement-to-sound generation but limits expressivity and complexity of interaction [10]. Supervised machine learning enables more complex movement-to-sound mappings but requires niche AI expertise, extensive time, and painstaking effort to customizably train [13]. Large unsupervised models such as diffusion can learn complex mappings automatically but require large amounts of data and computational power to train, are too heavy to customize for individual users, and are too slow for real-time interaction [17].

Our system seeks to address these gaps. It generates complex mappings between movement and sound spaces, customized to a user's data, using minimal resources and labor. It requires no hardware beyond a user's laptop, processing movement data from a laptop's camera via computer vision and training lightweight autoencoders to map between movement and audio spaces in real-time. Thus our system may reduce barriers to entry compared with previous state-of-the-art movement-to-sound paradigms.

# 2 Background & Motivation

Dancers have experimented with incorporating movement sonification into their practice for years [3]. They can map their body position to synthesizer parameters to control the generated sound [10]. With traditional synthesizers, this generally yields relatively direct and simple mappings between movement and sound. Neural audio generators which have emerged in the last several years may offer a more complex solution [5]. They can learn a latent audio space, with sound dimensions abstracted into a smaller, continuous, and more expressive set of parameters. By manipulating this latent audio space, artists can control more complex aspects of the generated sound.

Neural audio generators create sound by predicting the next section of a waveform. WaveNet [14] models the long temporal relationships inherent to music by incorporating dilated causal convolutions, similar to CNNs. Engel et al (2017) [7] applied WaveNet's approach to an autoencoder, enabling techniques such as timbre transfer. Caillon and Esling (2021) [5] broke autoencoder training into an initial representation learning stage, and a secondary adversarial fine-tuning stage. Four years later their open-source, user-friendly software "RAVE" remains standard in state-of-the-art audio performance and machine learning research [4] [13] [20]. Numerous models and variations are available online.

Researchers are exploring how artists can control neural audio generators from the human-performance space through embodied practice. Vigliensoni and Fiebrink (2023) [19] assigned RAVE's latent space subjective labels, which could be arranged in a 2D interface and manipulated interactively. Zheng et al (2024) [20] embedded real-time sketches into RAVE's latent space to be decoded as audio. Nabi et al (2024) [13] used a wearable motion sensor to map complex gestures (i.e. dance) to RAVE's latent space through three different movement-to-sound mapping methods, including training a Hidden Markov Regression model through supervised learning to map movement to RAVE's latent audio space. MM-Diffusion [17] models dance and other visual data jointly with audio, using multimodal diffusion to learn a shared audio-visual embedding space.

But current approaches for mapping embodied practice to a latent audio space for interactive exploration require specialized hardware, niche AI expertise, and painstaking labor and time to train [13] [20]. Many artists may find it hard to access these, particularly if they are new to the technology and unsure how much time, effort, and money they want to invest in it. Computer vision and unsupervised learning may help lower the barriers to entry, democratizing access for all.

Human bodies can be modeled from video using pose estimation, a computer vision task [11] [6] [1] [15]. Pose estimators identify relative locations in space of particular human body parts, e.g. joints (these are called "keypoints").

These keypoints locations are continuous <x, y> vectors which can be mapped to synthesizer parameters, including to a neural audio generator's latent space [13]. By constructing a mapping function between pose keypoints and synthesizer parameters, synthesizer output can be controlled via human body position. Our project investigates ways to construct these mappings.

# 3 System Overview

Human keypoints are extracted from video image data using pre-trained pose estimators (e.g. MoveNet). An autoencoder is trained on the user's custom movements to project the keypoints into a latent movement space. A multimodal encoder-decoder network may be trained on custom movement and audio data to project movement embeddings into audio space, or the multimodal network may be omitted and direct mapping used instead. The audio space mapped to is that of a pre-trained neural audio generator (e.g. RAVE). The audio generator can finally decode the embeddings to sound.
Our system relies on two pre-trained ML models – a pose estimator which extracts keypoints coordinates (e.g. MoveNet), and an audio generator which generates audio from a latent space (e.g. RAVE).
We train two additional autoencoders through unsupervised learning:

1. "Keypoints-Movement Projector" learns to encode keypoints coordinates (extracted from a pose estimator) into a "movement embedding" of the same dimensionality as the latent space of a pre-trained neural audio model (e.g. RAVE).
2. "Multimodal Movement-Audio Projector" learns to encode both movement embeddings and audio embeddings into a latent movement-audio space. This step can optionally be omitted, with movement embeddings mapped directly into latent audio space instead. Omitting this step saves time at training and reduces latency at inference, at the potential cost of mapping robustness, complexity, and customization to curated audio files.

Our system is generalizable and can accommodate any latent projector architecture or training procedure. For proof-of-concept, we construct each encoder and decoder as two linear layers of 4 to 34 neurons connected via ReLU activation, utilizing Adam optimization and mean squared error loss. We train these custom autoencoders, on ~30 seconds of custom movement data, for less than two minutes each. Thus our system requires only 30 seconds of the user's time to record movement data, and less than 5 minutes of total time to train. Once autoencoders are trained, users can interactively manipulate the learned movement-audio mappings. They can explore the latent audio space, translating their embodied practice to sound.

# 4 Discussion

Our novel system enables users to map and interactively sonify their movements without specialized hardware or niche AI expertise, centering humans in the generative process and improving accessibility and democratization of AI access over previous state-of-the-art movement-to-sound paradigms. Our system provides a direct possibility for users to sonify their movement using deep learning models. Our project expands the universe of available artistic expression, offering a new flavor of creation. It counters the threat of generative AI replacing people by incorporating the human role intimately within this new artistic medium.
Our system generalizes to mappings between any latent spaces. We are using it to map movement to audio, but the same principles can apply just as easily to any other pair of modalities. Thus we contribute not only an accessible system for sonification of embodied practice, but a new unsupervised paradigm for mapping between latent spaces.

# 5 Future Works

We are currently exploring semi-supervised interactive machine learning, incorporating a secondary supervised/interactive fine-tuning process [8] after the primary unsupervised training to give users more control over the latent mappings. We are also investigating extraction of human keypoints through an ensemble of pose estimators, to encode more expressive human features in latent movement embeddings; as well as techniques to transform latent space to maximize novelty of generated output. Our system is still under active development. It will ultimately be generalized to support a broad user base, and open-sourced.

In future works we will analyze, refine, and evaluate our system, including through user studies. We will employ technology probes to investigate how artists from different backgrounds (including dance, music, and film) interact with the system; and we will run user experiments to assess the utility, accessibility, and value of the system to a broad, inclusive pool of participants. We will define and apply metrics for evaluation to compare our system to current state-of-the-art approaches.

# 6 Conclusion

We have presented a new technique and system implementation for sonification of embodied practice. We use unsupervised learning to map multimodally between movement and audio spaces. Our method may be more accessible and easier to use than previous state-of-the-art approaches. Our system requires minimal labor, time, niche expertise, specialized hardware, and money access compared with previous state-of-the-art movement-to-sound paradigms. It requires only 30 seconds of a user's movements, recorded through their laptop camera; and a few minutes of compute time on their laptop. Then it is customized to their gestures and curated audio library, ready to use any time.
We have also outlined our next steps. We specified technical advancements we are investigating including semi-supervised interactive machine learning and ensemble pose estimation, and we clarified our methodology for evaluation.


**Figure 1: Overview of proposed system architecture.**

## REFERENCES

<bib id="bib1"><number>[1]</number>Rishabh Bajpai and Deepak Joshi. 2021. MoveNet: A deep neural network for joint profile prediction across variable walking speeds and slopes. https://ieeexplore.ieee.org/document/9406043 </bib>
<bib id="bib2"><number>[2]</number>Tove Bang and Sarah Alaoui. 2023. Suspended Circles: Soma Designing a Musical Instrument. https://dl.acm.org/doi/pdf/10.1145/3544548.3581488 </bib>
<bib id="bib3"><number>[3]</number>Frédéric Bevilacqua, Norbert Schnell, and Sarah Fdili Alaoui. 2011. Gesture Capture: Paradigms in Interactive Music/Dance System. https://www.researchgate.net/publication/267946741_Gesture_Capture_Paradigms_in_Interactive_Music_Dance_Systems </bib>
<bib id="bib4"><number>[4]</number>Nick Bryan-Kinns and Zijin Li. 2024. Reducing Barriers to the Use of Marginalised Music Genres in AI. In Proceedings of Explainable AI for the Arts Workshop 2024 (XAIxArts 2024) at ACM Creativity and Cognition 2023. https://arxiv.org/pdf/2407.13439</bib>
<bib id="bib5"><number>[5]</number>Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. https://arxiv.org/pdf/2111.05011.pdf </bib>
<bib id="bib6"><number>[6]</number>Zhe Cao et al. 2019. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. https://arxiv.org/pdf/1812.08008.pdf </bib>
<bib id="bib7"><number>[7]</number>Jesse Engel et al. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In International Conference on Machine Learning (pp. 1068-1077). https://arxiv.org/pdf/1704.01279 </bib>
<bib id="bib8"><number>[8]</number>Rebecca Fiebrink et al. 2009. A Meta-Instrument for Interactive, On-the-fly Machine Learning. In New Interfaces for Musical Expression 2009. https://www.cs.princeton.edu/sound/publications/FiebrinkTruemanCook_NIME2009.pdf </bib>
<bib id="bib9"><number>[9]</number>Jules Françoise, Sarah Fdili Alaoui, and Yves Candau. 2022. CO/DA: Live-Coding Movement-Sound Interactions for Dance Improvisation. In CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA. ACM, New York, NY, USA 13 Pages. https://doi.org/10.1145/3491102.3501916 </bib>
<bib id="bib10"><number>[10]</number>Andy Hunt and Marcelo Wanderley. 2002. Mapping performer parameters to synthesis engines. https://dl.acm.org/doi/abs/10.1017/S1355771802002030 </bib>
<bib id="bib11"><number>[11]</number>BeomJun Jo and SeongKi Kim. 2022. Comparative Analysis of OpenPose, PoseNet, and MoveNet Models for Pose Estimation in Mobile Devices. https://www.iieta.org/journals/ts/paper/10.18280/ts.390111 </bib>
<bib id="bib12"><number>[12]</number>Thomas Mitchell et al. 2012. Musical Interaction with Hand Posture and Orientation: A Toolbox of Gestural Control Mechanisms. https://www.nime.org/proceedings/2012/nime2012_272.pdf </bib>

<bib id="bib13"><number>[13]</number>Sarah Nabi et al. 2024. Embodied exploration of deep latent spaces in interactive dance-music performance. https://dl.acm.org/doi/pdf/10.1145/3658852.3659072 </bib>
<bib id="bib14"><number>[14]</number>Aäron van den Oord et al. 2016. Wavenet: A generative model for raw audio. https://arxiv.org/pdf/1609.03499.pdf </bib>
<bib id="bib15"><number>[15]</number>George Papandreou et al. 2018. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. https://arxiv.org/pdf/1803.08225 </bib>
<bib id="bib16"><number>[16]</number>Mitchel Resnick et al. 2005. Design Principles for Tools to Support Creative Thinking. https://doi.org/10.1184/R1/6621917.v1 </bib>
<bib id="bib17"><number>[17]</number>Ludan Ruan et al. 2023. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10219-10228).</bib>
<bib id="bib18"><number>[18]</number>Stefania Serafin et al. 2014. Controlling Physically Based Virtual Musical Instruments Using The Gloves. https://www.nime.org/proceedings/2014/nime2014_307.pdf </bib>
<bib id="bib19"><number>[19]</number>Gabriel Vigliensoni and Rebecca Fiebrink. 2023. Steering latent audio models through interactive machine learning. https://ualresearchonline.arts.ac.uk/id/eprint/20199/1/VigliensoniFiebrink_ICCC2023.pdf </bib>
<bib id="bib20"><number>[20]</number>Shuoyang Zheng et al. 2024. A Mapping Strategy for Interacting with Latent Audio Synthesis Using Artistic Materials. https://arxiv.org/pdf/2407.04379 </bib>