GestureEncoder: a Joint Embedding Model for Movement-Based Musical Performance

JASON BRENT SMITH, Northwestern University, USA

BRYAN PARDO, Northwestern University, USA

Artificial Intelligence has been used to great effect to support musicians in live performance through gesture recognition, enabling control of the output of interactive music systems using movement. However, musicians working with gesture recognition must often manually map each new gesture to specific musical parameters, a lengthy and constrained process that is usually infeasible for live improvisation. The inner workings of AI systems are hidden from users, making it difficult for them to understand if the system's mappings between gestures and music align with their own. This contrasts with pairs of human improvisers who can build a shared understanding of mappings from motion to music in real time. This paper presents *GestureEncoder*, a model that maps gestures to musical outcomes in a way grounded in human understanding: a user can define gestures alongside a verbal description, and the system generates a joint embedding space of positional gesture data and user-defined physical and musical descriptions. *GestureEncoder* enables a user to quickly and organically create a set of motion-based musical instructions for human-understandable human-machine improvisation, using a two-dimensional visualization to represent how gestures are mapped to physical and musical descriptions.

Additional Key Words and Phrases: Artificial Intelligence, Music, Explainable AI

ACM Reference Format:
Jason Brent Smith and Bryan Pardo. 2026. GestureEncoder: a Joint Embedding Model for Movement-Based Musical Performance. *J. ACM* 37, 4, Article 111 (August 2026), 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In a live improvised musical performance, musicians who perform together collaborate to define a shared musical language built from gestures and their meanings. These gestures can take simple yet subjective forms, such as a band leader lifting and lowering their head to subtly signal to the rest of the musicians that a new section of a song is about to begin. These languages can be given more formal or regimented definitions: for example, *Conduction* is a trademarked (protocol/language) by Lawrence D. "Butch" Morris that relies on the live enactment of motion gestures by an improviser's orchestra leader, which are then refined and iterated on to enable the group to act as a unit [26].

Generative Artificial Intelligence (AI) is a rapidly growing area of research that has been incorporated into many music production workflows, with standard inputs to these systems often being text describing entire musical passages to generate [3, 19]. Interactive machine learning has also been used to map gestures to musical parameters during live musical performances, leading to the development of gesture-controlled music interfaces and the study of the affordances of the machine learning process itself [10]. For example, Google's MediaPipe [29] platform can be integrated into live camera applications for music performance, using classifiers for hand signs (e.g., open hand) and motions

Authors' Contact Information: Jason Brent Smith, jason.smith1@northwestern.edu, Northwestern University, Evanston, Illinois, USA; Bryan Pardo, pardo@northwestern.edu, Northwestern University, Evanston, Illinois, USA.

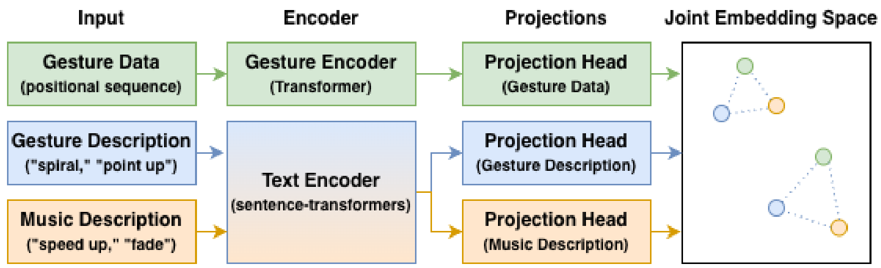

Fig. 1. *GestureEncoder* architecture. Gesture data sequences are encoded using a Transformer model, along with a language model for the gesture and music description labels. Three projection heads are combined to form a shared embedding space.

(e.g., circular pattern) to create low-dimensional representations of gestures that can be linked to audio parameters [23–25]. Although these motion-controlled systems are more capable of real-time performances, their "black-box" nature has limited performers' and audiences' ability to understand how the system *interprets* gestures: explainability and understanding the mechanisms of an AI-based music performance system is critical to allowing artists to control the creative process and providing audiences "emotional and psychological insight" into the performance [2]. This work explores how real-time, human-AI musical improvisation can be made more understandable by enabling the AI to connect a user's performance gestures to their musical intent.

We present *GestureEncoder*, a model for a human-understandable connection between physical motion and musical intent for improvisation on a level that mirrors Morris' *Conduction.* This model encodes low-dimensional representations of gesture data alongside user-defined labels for that gesture's meaning (both on a literal, physical level as well as the gesture's intended musical output) into a shared embedding space [9]. Our goal is to use shared embeddings so that a generative model traditionally controlled by text can be controlled by gestures. This approach provides an explainable real-time connection between motion and sound through a shared understanding of both modalities.

Embedding spaces are organizational structures in which topics can be arranged according to distances based on similarity [9], and various inputs have been used to generate conditional audio using embedding spaces. These spaces provide a method of applying explainable AI principles to user-controlled generative models [4] due to the understandable spatial relationship between visualized embedding spaces aligning with a musician's use of the space to control generative output. Shared embedding spaces have been used to connect music and language for the generation of audio and symbolic music based on text input [12, 15, 17], timbre descriptors [22], as well as controllable latent representations of audio through the RAVE family of autoencoder models [5–7]. MusicLM uses three models for audio and text representation and allows descriptive text to control generated audio [1]. VampNet generates variations of a given input audio segment by modeling masked acoustic tokens [11]. Diffusion models such as Stable Audio Open Small [18] and Magenta Realtime [27] enable real-time text-controlled audio generation. Using a shared gesture-sound latent space, *GestureEncoder* generates human-understandable prompts for language models, taking as input a user's motions and creating a "vocabulary" that represents the intention of those motions. Unlike prior systems by the authors [24], which allow the user to indirectly map labeled gestures to unknown audio parameters using a classifier, followed by user feedback to alter those mappings, *GestureEncoder* allows the user to define and make real-time adjustments of gesture-music mappings to direct instructions for diffusion model-style output. Additionally, *GestureEncoder* organizes learned gesture-music combinations in a human-understandable shared embedding space for applications such as zero-shot retrieval or interpolation between gestures.

2 *GestureEncoder*

GestureEncoder is a joint embedding model developed in PyTorch that combines three modalities:

- **Gesture Data:** Sequences of positional data captured through the Google MediaPipe hand-tracking model [29].
- **Gesture Description:** Verbal descriptions of physical motion (e.g., "spiral", "point up").
- **Music Description:** An intended musical outcome of a gesture (e.g., "speed up", "fade").

The user can record gestures and label them with motion descriptors and intended musical outcomes, thereby creating a data set of triplets: gesture, text, and music. Gestures are encoded using PyTorch `nn.TransformerEncoderLayer` and `nn.Transformer` layers with 256 dimensions. To encode the gesture description and music description labels, we use frozen, pre-trained Sentence Transformers[1] models [20] based on BERT architectures [14]. All three representations use a projection head with `nn.ReLU` and linear activation functions to form a 128-dimensional embedding. During training, *GestureEncoder* uses contrastive loss [13] by combining a PyTorch implementation[2] of InfoNCELoss [21] across the three domains through a weighted sum of InfoNCELoss (L) for each combination of gesture data (g), gesture description (d), and music description (m). Comparisons of g or d against m were weighted by $\alpha(0.75)$ and $\beta(0.75)$, relative to the comparison of g and d, in order to maximize the connection between gestures and their physical descriptions while enabling more uncertainty in their connections to musical outputs.

$$L = L_{g,d} + \alpha L_{g,m} + \beta L_{d,m} \tag{1}$$

3 Evaluation

We use a very limited data set of five gestures (Table 1) to demonstrate a feasible set of instructions that could be presented to *GestureEncoder* and labeled in one minute in a hypothetical performance situation. The average training time for these five gestures is *75.95* seconds over 300 epochs on an Apple M2 MacBook Pro 2022, which represents a small portion of the up to three-hour sessions of teaching “vocabulary” to musicians by Butch Morris when recording performances using *Conduction* [8].

Gesture Data	Gesture Description	Music Description
10 sec. video	Wave Hand	Speed Up
8 sec. video	Clap Hands	Stop
10 sec. video	Thumbs Up	Louder
10 sec. video	Point Left	Trombone
10 sec. video	Point Right	Trumpet

Table 1. Five user-defined, *Conduction*-style gestures.

Figure 2 shows a two-dimensional uniform manifold approximation and projection (UMAP) [16] visualization of the trained *GestureEncoder*, using cosine similarity. Each “triangle” represents a triplet of positional data, Gesture Description, and Music Description, with similar gestures grouped together. Each item is closer to the other members of its “triangle” than to any other item; as a result, on this small dataset, the retrieval accuracy and mean reciprocal rank (MRR) are both 100% for each combination of the three modalities, indicating that each gesture/music description item

[1] https://pypi.org/project/sentence-transformers/
[2] https://github.com/RElbers/info-nce-pytorch

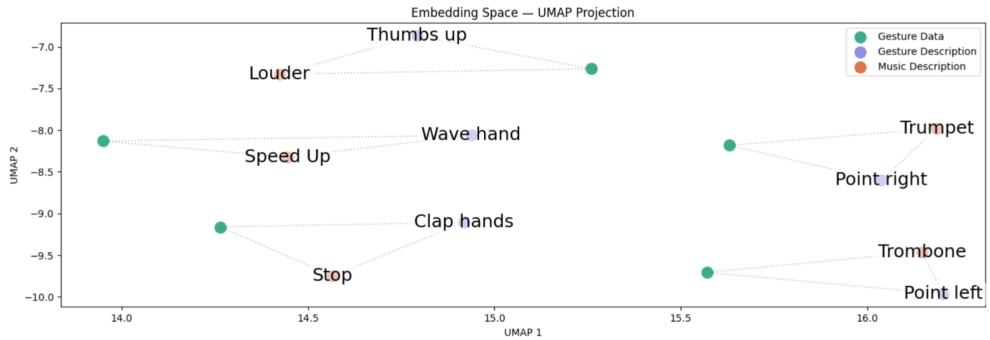

Fig. 2. The shared embedding space of *GestureEncoder* represented in two UMAP dimensions. For each gestural information (green), an equivalent motion descriptor (blue) and musical meaning (orange) are connected by a dotted line.

is the "first rank", or closer to its corresponding gesture data item than all other descriptions. Gestures with an inherent similarity, such as how "point right" and "point left" are closer to each other than they are to the "thumbs up" gesture.

Geometric alignment (the average cosine similarity between pairs) and uniformity (the log of the average Gaussian kernel between pairs) are two metrics used to assess the quality of a representation learned with contrastive loss [28]. The alignment of each pair *GestureEncoder's* modalities is 0.883 (gesture data, gesture description), 0.883 (gesture data, music description), and 0.926 (gesture description, music description), with each pair having a uniformity of -4.6. These results indicate greater alignment between the two text representations, as expected, since both were trained with the same encoder. Equal scores between the gesture and the two text modalities and the overall negative uniformity indicate an even spread of the "triplets" across the shared embedding space.

4 Conclusion

GestureEncoder is a prototype joint-embedding model that supports human-AI musical improvisation by enabling a user to define a "vocabulary" of physical movements alongside their intended musical outputs, allowing them to use gestures to control a language model in live performance settings. A user can define a set of performance gestures using a few video examples, along with labels indicating the gesture's semantic meaning and the intended musical outcome. As a result, *GestureEncoder* can recognize performance gestures and compare their meanings in a semantic space that represents similarities between gestures and musical outcomes. In future work, we will expand on this by interpolating between known musical meanings to generate language-model-friendly musical instructions for previously unseen, undefined gestures. We also aim to incorporate *GestureEncoder* into live musical applications, leveraging its ability to map motion to human-understandable musical descriptions to serve as live prompting input for a generative audio model, while visualizing the shared embedding space, and to observe how performers use and adapt to this tool.

Acknowledgments

This work was supported by NSF Award Number 2300633. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* (2023).

[2] Jack Armitage, Nicola Privato, Victor Shepardson, and Celeste Betancur Gutierrez. 2023. Explainable AI in Music Performance: Case Studies from Live Coding and Sound Spatialisation. In *XAI in Action: Past, Present, and Future Applications*. https://openreview.net/forum?id=se4ojQqjB5

[3] Misagh Azimi and Mo H Zareei. 2025. Live Improvisation with Fine-Tuned Generative AI: A Musical Metacreation Approach. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 389–393.

[4] Nick Bryan-Kinns, Berker Banar, Corey Ford, Courtney N. Reed, Yixiao Zhang, Simon Colton, and Jack Armitage. 2023. Exploring XAI for the Arts: Explaining Latent Space in Generative Music. arXiv:2308.05496 [cs.AI] https://arxiv.org/abs/2308.05496

[5] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011* (2021).

[6] Franco Caspe, Andrew McPherson, and Mark Sandler. 2025. Waveform Autoencoding at the Edge of Perceivable Latency. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, Article 10, 4 pages. doi:10.5281/zenodo.15699550

[7] Franco Caspe, Jordie Shier, Mark Sandler, Charalampos Saitis, and Andrew McPherson. 2025. Designing Neural Synthesizers for Low Latency Interaction. *arXiv preprint arXiv:2503.11562* (2025).

[8] Farai Chideya. 2008. Butch Morris on the Art of 'Conduction'. NPR *News & Notes*. https://www.npr.org/transcripts/19145728 Accessed April 30, 2026.

[9] Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* 8 (2020), 439–453.

[10] 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*. 237–242.

[11] H Flores Garcia, P Seetharaman, R Kumar, and B Pardo. 2023. VampNet: Music Generation via Masked Acoustic Token Modeling. 24th International Society for Music Information Retrieval Conference.

[12] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415* (2022).

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

[14] Mikhail V Koroteev. 2021. BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943* (2021).

[15] Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. 2024. Efficient neural music generation. *Advances in Neural Information Processing Systems* 36 (2024).

[16] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[17] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. 2021. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091* (2021).

[18] Zachary Novack, Zach Evans, Zack Zukowski, Josiah Taylor, CJ Carr, Julian Parker, Adnan Al-Sinan, Gian Marco Iodice, Julian McAuley, Taylor Berg-Kirkpatrick, et al. 2025. Fast Text-to-Audio Generation with Adversarial Post-Training. *arXiv preprint arXiv:2505.08175* (2025).

[19] Y Yogi Tegar Nugroho and P Paulus Metta Dwi Manggala. 2024. The use of AI in creating music compositions: A case study on Suno application. In *7th Celt International Conference (CIC 2024)*. Atlantis Press, 177–189.

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[21] Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S Zimmermann, and Wieland Brendel. 2024. Infonce: Identifying the gap between theory and practice. *arXiv preprint arXiv:2407.00143* (2024).

[22] Jordie Shier, Rodrigo Constanzo, Charalampos Saitis, Andrew Robertson, and Andrew McPherson. 2025. Designing Percussive Timbre Remappings: Negotiating Audio Representations and Evolving Parameter Spaces. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, Article 66, 10 pages. doi:10.5281/zenodo.15698926

[23] Jason Smith and Jason Freeman. 2021. Effects of Deep Neural Networks on the Perceived Creative Autonomy of a Generative Musical System. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 17. 91–98.

[24] Jason Smith and Jason Freeman. 2025. Adaptation and Perceived Creative Autonomy in Gesture-Controlled Interactive Music. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, Article 55, 10 pages. doi:10.5281/zenodo.15698904

[25] Jason Brent Smith and Jason Freeman. 2023. Effects of Visual Explanation on Perceived Creative Autonomy in an AI-Based Generative Music System. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 25–28.

[26] Thomas Taylor Stanley. 2009. Butch Morris and the Art of Conduction. (2009).

[27] Lyria Team, Antoine Caillon, Brian McWilliams, Cassie Tarakajian, Ian Simon, Ilaria Manco, Jesse Engel, Noah Constant, Yunpeng Li, Timo I Denk, et al. 2025. Live music models. *arXiv preprint arXiv:2508.04651* (2025).

[28] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*. PMLR, 9929–9939.

[29] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* (2020).