arts

MDPI

Review

Embodied Musical Gesture Analysis from Audio Representations: A Review of Datasets, Audio Descriptors, and Deep Learning Approaches

Ioanna Miliaresi

Department of Audio and Visual Arts, Ionian University, Tsirigoti Sq. 7, 49100 Corfu, Greece; imiliaresi@ionio.gr; Tel.: +30-261086373

Abstract

Machine learning approaches have enabled increasingly powerful analysis of musical audio, allowing computational models to capture complex acoustic patterns related to musical structure, performance, and expression. However, whether learned audio representations preserve information related to the bodily actions underlying musical sound production remains an open research question. This review examines how machine-learned audio representations encode information associated with embodied musical processes by connecting perspectives from embodied music cognition, computational music analysis, and multimodal machine learning. It establishes a conceptual framework that organizes embodied musical information according to gesture type, acoustic observability, and the relationship between physical action and sound production, distinguishing sound-producing gestures, expressive movements, individual embodiment, and collective performance processes. The review then surveys datasets for audio–gesture research, including motion-capture recordings, instrument-performance datasets, audiovisual resources, and large-scale audio-only datasets, highlighting challenges related to embodiment fidelity, scalability, annotation, and multimodal alignment. It further examines conventional audio descriptors, deep neural and self-supervised audio representations, cross-modal learning approaches, and evaluation methodologies for assessing embodied information within learned representations. Particular attention is given to gesture–sound ambiguity, limited benchmarking resources, generalization across performers and instruments, and the distinction between gesture-related encoding and acoustic correlation. Current evidence suggests that learned audio representations contain partial and task-dependent traces of embodied musical information, particularly for gestures with strong acoustic coupling, while evidence for higher-level expressive and communicative movements remains limited. Determining whether these representations capture meaningful relationships between physical performance processes and musical sound remains an open challenge requiring multimodal, interpretable, and perceptually grounded evaluation frameworks.

Keywords: deep learning; musical gesture analysis; audio descriptors; music embodiment

Check for updates

Academic Editor: Stella Paschalidou

Received: 14 July 2026
Revised: 16 September 2026
Accepted: 18 September 2026
Published: 22 September 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Embodied music cognition conceptualizes musical experience as emerging from corporeally mediated interactions rather than passive auditory perception. Within this framework, musical meaning arises through the dynamic coupling of body, sound, and environment, whereby perception and action are continuously intertwined. Godøy proposed that

listeners internally simulate sound-producing actions, suggesting that movement-related representations contribute to music perception (Godøy 2003). Leman further characterized the body as a mediator between sound production and perceptual experience through action–perception coupling (Leman 2007). Subsequent research has expanded this perspective by incorporating sensorimotor grounding and predictive mechanisms linking movement and auditory perception (Maes et al. 2014).

Within this theoretical perspective, musical gestures are understood as meaningful movement-related processes that connect physical action, sound production, and expressive interpretation. They encompass both sound-producing gestures that directly generate acoustic events and ancillary expressive movements that facilitate communication during performance without directly influencing sound production (Godøy 2003; Jensenius et al. 2009). This distinction is particularly important for computational analysis because different gesture types exhibit varying degrees of acoustic observability. Instrumental actions often leave measurable traces in timbre, timing, and dynamics, whereas expressive body movements may be only weakly or indirectly reflected in the audio signal.

Early computational studies investigated musical gesture through multimodal analyses integrating movement, sound, and performer interaction. Camurri, Leman, and colleagues established the foundations for computational modelling of embodied musical performance by developing integrated audio–movement frameworks (Camurri and Volpe 2011; Camurri et al. 2003; Leman and Camurri 2006). Subsequent research extended these approaches to gesture recognition, gesture–sound mapping, and interactive digital musical systems using motion capture, video, and sensor-based measurements (Caramiaux and Tanaka 2013; Caramiaux et al. 2014).

In parallel, advances in music information retrieval (MIR) and deep learning have led to powerful learned audio representations, including convolutional, recurrent, transformer-based, and self-supervised models that capture complex temporal, spectral, and contextual structures in musical audio (Baevski et al. 2020; Cramer et al. 2021; Liang 2023). Unlike traditional gesture-analysis approaches, these models are typically trained solely on audio and therefore lack direct access to the physical actions underlying sound production. Their success in performance analysis, expressive performance modelling, and general audio understanding raises a fundamental question: when trained primarily on audio, do learned representations retain information related to the bodily processes underlying musical sound production?

From an artificial intelligence perspective, this question can be formulated as an inverse inference problem: determining whether latent audio representations contain information related to the physical actions that give rise to musical sound. Throughout this review, *encoding* refers to the extent to which performer gesture information can be inferred from learned audio representations, either explicitly through downstream prediction tasks or implicitly through representation probing. Importantly, encoding does not imply that a model has acquired a causal or physically grounded understanding of gesture; rather, it indicates that gesture-related information is present and accessible within the learned representation space.

Inferring gesture from audio is inherently challenging because audio contains only the acoustic consequences of bodily action rather than the actions themselves. Different bodily configurations may produce acoustically similar outcomes, while similar gestures can generate different sounds depending on the instrument, performer, expressive intention, and performance context (Maes 2016). Furthermore, expressive or ancillary gestures may not contribute directly to sound production and therefore leave only weak or ambiguous traces in audio representations. These challenges are also reflected in the available datasets. Multimodal corpora provide rich measurements of embodied performance but remain

relatively limited in scale, whereas large-scale audio-only datasets facilitate representation learning at scale while capturing embodiment only implicitly.

Despite extensive research in embodied music cognition, expressive performance modelling, and machine learning for music, the relationship between embodied musical gestures and learned audio representations remains insufficiently understood. Existing reviews have examined sensorimotor grounding in music perception (Maes 2016), computational models of expressive performance (Cancino-Chacón et al. 2018), embodied perspectives on musical knowledge (Reybrouck and Schiavio 2024), and machine learning approaches in music research more broadly (Liang 2023). However, none directly addresses whether contemporary audio representation-learning methods encode, preserve, or discard information related to performer gestures.

Unlike previous reviews that separately consider embodied music cognition, expressive performance modelling, or audio representation learning, this review focuses specifically on the extent to which learned audio representations preserve information related to the embodied processes underlying musical performance.

This review follows a structured narrative approach to examine whether machine-learned audio representations preserve information related to embodied musical processes. The literature was identified through searches of major academic databases, including Google Scholar and Scopus, using combinations of keywords related to embodied music cognition, musical gesture, audio–gesture datasets, audio features for music gesture analysis, audio representation learning, deep learning for music analysis, cross-modal learning, and music information retrieval. Approximately 100 peer-reviewed journal articles and conference papers were initially collected and examined. The identified literature was subsequently screened and refined according to its relevance to the central research question of this review. Studies were included when they contributed to one or more of the following themes: (i) theoretical foundations of embodied music cognition and musical gesture, (ii) datasets linking audio with performer movement or expressive behaviour, (iii) acoustic descriptors and learned audio representations for music analysis, (iv) deep learning and cross-modal approaches for audio–gesture modelling, and (v) evaluation methodologies for assessing gesture-related information in learned representations. The selected studies were then organized thematically rather than chronologically to establish connections between theoretical concepts, computational methods, available resources, and current limitations. Seminal works were included to define foundational concepts and methodological developments, while recent studies were emphasized to reflect advances in deep learning, self-supervised representation learning, and multimodal modelling. The resulting synthesis aims to identify current evidence, methodological challenges, and future directions for determining whether learned audio representations capture meaningful embodied information rather than only acoustic correlations (Figure 1).

This review addresses this gap by critically examining whether and how machine-learned audio representations encode embodied gestural information underlying musical performance. It integrates perspectives from embodied music cognition, computational music analysis, and multimodal learning to investigate whether physical actions, expressive movements, and performer interactions leave measurable or inferable information within learned audio representations. The review first establishes a theoretical and computational framework for musical gesture by defining different levels of embodiment, acoustic observability, and the challenges associated with inferring physical action from acoustic information (Section 2). It then critically examines datasets for audio–gesture analysis, including motion-capture, audiovisual, symbolic performance, and large-scale audio-only resources, highlighting the trade-offs among embodiment fidelity, scalability, annotation availability, and the extent to which embodied information is explicitly

represented (Section 3). Next, it reviews conventional acoustic descriptors and modern representation-learning methods, including deep neural networks and self-supervised models, with particular attention to their capacity to preserve gesture-related information. The audio descriptors are also formally described through mathematical formulations of commonly used features, including spectrograms, Mel-frequency representations, MFCCs, RMS amplitude, spectral centroid, spectral flux, fundamental frequency, and onset-related features (Section 4). The review subsequently examines machine learning approaches for modelling audio–gesture relationships, including deep learning architectures and cross-modal learning strategies, with emphasis on how these approaches learn, preserve, or infer gesture-related information from acoustic data (Section 5). Finally, it examines evaluation methodologies, current limitations, and future research directions aimed at determining whether learned audio representations capture meaningful embodied structures rather than merely reflecting acoustic correlations (Section 6). The main findings, research challenges, and future research directions are summarized in the conclusion (Section 8).

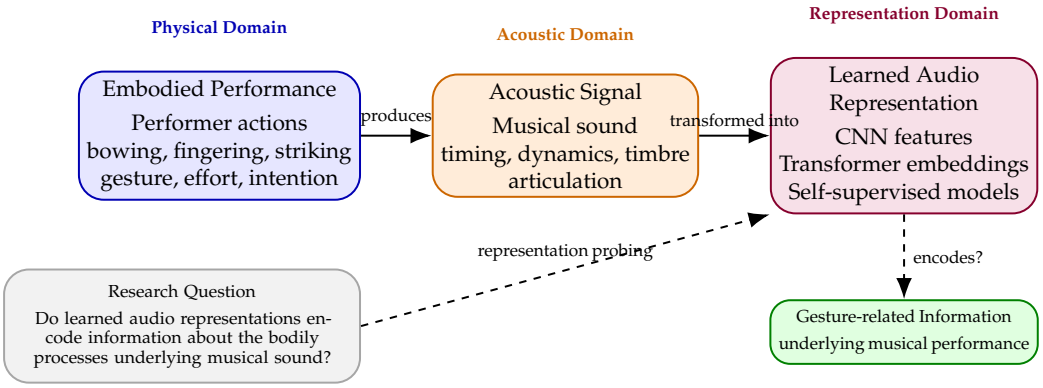


Figure 1. Conceptual framework linking embodied musical performance and learned audio representations. Performer actions generate acoustic consequences that are transformed into learned audio representations. The framework illustrates the central question of whether these representations encode information related to the gestures underlying musical performance.

2. Embodied Music and Gesture Framework

This section establishes a theoretical and computational perspective on musical gesture within embodied music cognition and machine learning. Rather than considering gesture as either purely physical movement or a perceptual construct, we conceptualize it as a multi-level phenomenon linking bodily action, sound production, expressive intention, and their acoustic and multimodal consequences. This perspective motivates the investigation of whether learned audio representations encode information related to embodied musical processes.

2.1. Defining Musical Gesture in Embodied Music Cognition

Embodied music cognition proposes that musical experience emerges through continuous interactions among bodily action, perception, and the surrounding environment. From this perspective, sound is not an isolated auditory object, but rather the acoustic outcome of interactions among performers, instruments, and listeners. Godøy proposed that listeners internally simulate sound-producing actions, suggesting that movement-related representations contribute to music perception (Godøy 2003). Leman described musical experience as arising through action–perception coupling, whereby bodily engagement mediates the relationship between sound production and interpretation (Leman 2007). Subsequent studies have highlighted the role of sensorimotor learning and predictive mechanisms in linking movement and auditory perception (Maes et al. 2014).

From a computational perspective, musical gesture can be understood as an unobserved physical process that produces observable traces in sound. Audio does not directly represent bodily movement; rather, it captures the acoustic consequences of performer actions. Consequently, learned audio representations may encode gesture-related information without explicitly modelling movement, raising the question of whether artificial intelligence systems capture aspects of embodied musical performance.

Musical gestures can be broadly divided into two categories:

- Sound-producing gestures: physical actions that directly generate acoustic events, such as bowing, striking, fingering, or blowing. These gestures are closely related to acoustic properties including articulation, timing, dynamics, and timbre.
- Ancillary expressive gestures: movements that do not directly generate sound but contribute to communication, coordination, and expressive interpretation, such as body sway, head movement, or conducting gestures. These gestures may contribute to performance perception while leaving weak or ambiguous traces in the acoustic signal.

2.2. Levels of Embodied Musical Information

Embodied musical information can be organized hierarchically according to its relationship with physical action and degree of acoustic observability. Lower levels correspond to processes more directly coupled with sound production, whereas higher levels involve increasingly abstract expressive and perceptual processes (Figure 2).

- Sound-producing gestures: motor actions directly involved in generating acoustic events, such as bowing, striking, fingering, and articulation. These exhibit the strongest acoustic relationship due to their direct role in sound production (Tindale et al. 2011; Traube et al. 2003).
- Performance dynamics: motor control processes shaping timing, articulation, dynamics, and micro-temporal variation. Although not necessarily observable as discrete gestures, they produce measurable acoustic patterns (MacRitchie et al. 2017).
- Ancillary expressive gestures: non-sound-producing movements involved in communication and interpretation, including posture, body motion, and conducting gestures. These affect perceived expressiveness while producing weaker acoustic correlates (Clayton et al. 2022; Godøy 2003; Nadkarni et al. 2023).
- Sensorimotor representations and perceived motion: internal representations of movement activated during listening, reflecting embodied interpretation rather than directly observable performer actions (Leman 2007; Maes et al. 2014; Schiavio et al. 2015).
- Expressive intention: higher-level interpretative and affective aspects related to phrasing, emotion, and performer intent. These influence acoustic outcomes but are indirectly reflected in acoustic features (Leman 2007; Reybrouck and Schiavio 2024).

This hierarchy reflects a continuum of acoustic observability: some aspects can be measured directly through sensors, whereas others remain latent and require inference from acoustic or multimodal data.

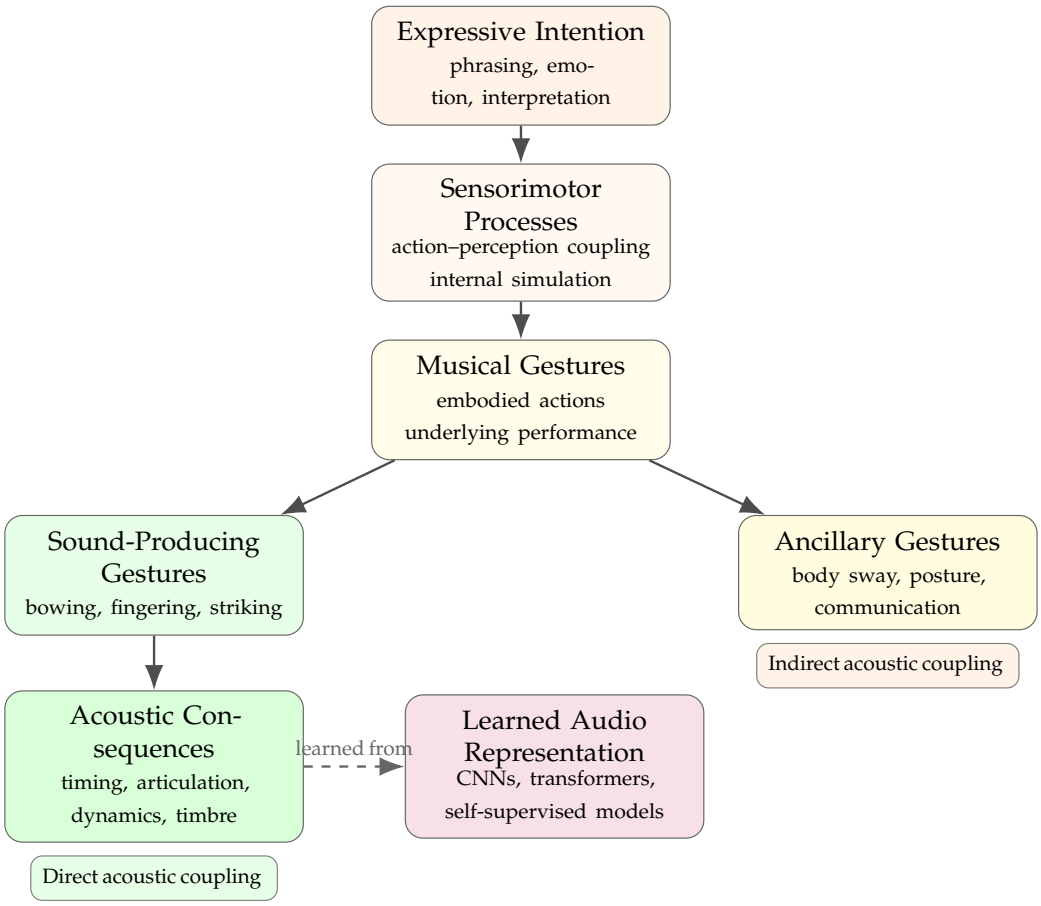


Figure 2. Hierarchy of embodied musical processes organized according to their relationship with physical action and acoustic observability. Performer intentions and sensorimotor processes give rise to musical gestures, which differ in their relationship with sound production. Sound-producing gestures exhibit stronger acoustic coupling, whereas ancillary gestures are more indirectly reflected in the acoustic signal.

2.3. Individual and Collective Embodiment

Embodied musical information can also be distinguished by the social organization of performance.

- Individual embodiment: refers to the bodily actions and expressive strategies of a single performer, including instrumental technique, articulation, timing variation, and performer-specific movement patterns. These are commonly studied using motion capture, instrument sensors, and isolated recordings.
- Collective embodiment: refers to coordinated bodily processes emerging through interactions among performers, including synchronization, entrainment, leadership, and shared timing adaptation. These may be reflected in audio through timing relationships and ensemble dynamics but are difficult to attribute to individual movements.

This distinction is relevant for learned audio representations, as models trained on ensemble recordings may capture collective performance patterns without explicitly encoding individual gestures.

2.4. Observability and Representational Spaces of Gesture

From a computational perspective, musical gesture can be categorized by its degree of observability:

- Directly observable embodiment: captured through motion capture, inertial sensors, vision-based pose estimation, or instrument-embedded sensors, providing explicit movement measurements.
- Indirectly observable embodiment: inferred from acoustic signals, symbolic representations such as MIDI or scores, or performance descriptors including timing, dynamics, and spectral features.
- Latent embodied representations: unobserved structures learned by computational models, particularly deep neural networks trained on audio or multimodal data.

This framework connects embodied music cognition with representation learning: motion-based systems measure embodiment directly, whereas audio-only models infer gesture-related information from acoustic consequences.

Gesture information can be represented through complementary spaces, including continuous kinematic trajectories, discrete gesture categories, probabilistic latent variables, and joint audio–motion embeddings learned through contrastive or reconstruction objectives. These representations differ in interpretability and proximity to physical embodiment, ranging from explicit movement measurements to abstract latent spaces learned from large-scale audio data.

2.5. Computational Challenges in Gesture Encoding

The relationship between gesture and sound is inherently ambiguous. Similar acoustic outcomes may result from different bodily actions, while similar gestures may produce different sounds depending on instrument mechanics, performer characteristics, expressive intention, and context (Maes 2016). Therefore, gesture inference from audio is an ill-posed inverse problem due to the non-unique mapping between sound and movement.

A key challenge in representation learning is separating gesture-related information from correlated acoustic factors. A model may appear to encode gesture because movement affects sound, while instead relying on variables such as instrument identity, recording conditions, musical style, or performer characteristics. Assessing whether representations capture embodied information therefore requires carefully designed probing tasks and evaluation protocols.

Dataset limitations further complicate this problem. Motion-capture and multimodal datasets provide detailed embodiment measurements but remain limited in scale and diversity, whereas large-scale audio datasets enable representation learning at scale but offer only indirect access to the physical actions underlying sound production. This motivates multimodal approaches that use movement, video, or sensor data as supervision to investigate whether audio representations preserve gesture-related information (Caramiaux and Tanaka 2013; Françoise et al. 2013).

The framework presented above connects embodied music cognition with computational modeling by distinguishing forms of musical gesture according to acoustic observability and abstraction level. Building on this foundation, audio representations can be viewed as intermediate spaces that transform acoustic consequences of physical actions into computational features, which may retain information related to embodied musical processes (Figure 1). The following sections examine audio representations, datasets, learning paradigms, and evaluation methodologies used to assess whether these representations encode embodied musical information.

2.6. Practical Applications and Rationale of Musical Gesture Analysis

Musical gesture analysis has practical relevance because performer movements are closely connected to sound production, musical expression, and interaction with digital musical systems. Analysing these relationships can provide information that is not readily

available from audio or movement considered independently. In particular, gesture analysis can help characterize how physical actions contribute to changes in articulation, timbre, dynamics, and timing, thereby providing a computational perspective on the embodied nature of musical performance (Caramiaux et al. 2014; Traube et al. 2003).

One important application is the analysis and recognition of instrumental performance techniques. Audio and movement information can be used to identify gesture categories and to investigate how different physical actions produce distinguishable acoustic consequences. For example, machine learning approaches have been applied to the classification of violin bowing gestures, demonstrating that acoustic and movement-related information can support the recognition of different bowing patterns (Dalmazzo and Ramirez 2019). More recent deep learning approaches have further investigated the estimation of musical gesture patterns from synchronized audio and motion recordings (Dalmazzo and Ramirez 2021). Such applications are relevant not only for automatic gesture recognition, but also for understanding the relationship between physical performance actions and their resulting sound.

A second application concerns interactive music systems and digital musical instruments. Gesture analysis can be used to establish mappings between performer actions and sound parameters, allowing physical movement to control sound synthesis, selection, or transformation. Caramiaux et al. demonstrated several gesture–sound mapping scenarios in which temporal and acoustic characteristics of sound are related to corresponding movement patterns (Caramiaux et al. 2014). Similarly, machine-learning-based gesture recognition has been applied to interactive virtual musical instruments, illustrating how recognized hand movements can be used as control inputs for real-time musical interaction (Kritsis et al. 2018). These applications motivate the analysis of gesture because understanding the relationship between movement and sound is necessary for designing expressive and responsive interaction mechanisms.

Gesture analysis is also relevant to expressive performance modelling. Physical movements can provide complementary information about expressive intentions that may be difficult to infer from audio alone. Conducting gestures, for example, provide a means of communicating tempo, dynamics, and interpretative intentions, making their analysis relevant to systems that model or reproduce expressive musical control (Chen et al. 2018). More generally, analysing gesture together with audio can help distinguish acoustic variations caused by differences in performance actions from variations arising from other factors such as musical content or instrument characteristics.

The underlying rationale for musical gesture analysis is therefore not limited to improving recognition accuracy. Rather, gesture analysis provides a means of investigating the relationship between physical action and sonic outcome. This perspective is particularly important for embodied music research because the same acoustic event may result from different physical actions, while similar physical actions may produce different sounds depending on the performer, instrument, and performance context. Consequently, analysing gesture–sound relationships can contribute to performance analysis, interactive music technology, expressive modelling, and the development of representations that retain information about embodied musical processes (Caramiaux et al. 2014; Traube et al. 2003).

From this perspective, the practical applications also motivate the broader research question addressed in this review: whether machine-learned audio representations retain information related to the physical actions underlying musical sound. If such information is present and transferable across performers, instruments, and performance contexts, audio representations could provide a scalable means of studying aspects of musical embodiment even when direct movement measurements are unavailable.

3. Datasets for Audio–Gesture Representation Learning

Investigating whether machine-learned audio representations encode gesture-related information associated with embodied musical processes requires datasets linking sound with physical action. Unlike conventional music datasets designed for classification, retrieval, or generation, gesture-oriented datasets must capture evidence of the embodied processes underlying performance, enabling evaluation of whether audio representations encode information related to performer actions, expressive behavior, and movement structures.

Datasets differ in their degree of embodiment observability. Directly observable embodiment includes motion capture, inertial sensors, instrument sensors, and pose estimation, providing explicit measurements of physical actions. Indirectly observable embodiment includes audiovisual recordings, symbolic performance data, and expressive annotations that provide partial evidence of bodily behavior. Latent gesture-related information refers to movement-related patterns inferred from audio representations without explicit gesture supervision.

This distinction defines evaluation scope: motion-based datasets enable direct gesture prediction, whereas audio-only datasets reveal acoustic correlates associated with performance. Thus, embodiment fidelity and dataset scale remain in tension: multimodal datasets provide richer physical evidence but limited scale, while large-scale audio datasets support representation learning but capture embodiment only indirectly.

3.1. Direct Embodiment Datasets

Datasets containing synchronized audio and performer movement provide the most direct empirical basis for studying audio–gesture relationships. They enable evaluation of whether audio representations encode information related to the physical actions underlying sound production.

Instrument-performance datasets provide fine-grained gesture measurements. The MOSA dataset combines audio, high-resolution 3D motion capture, and semantic performance annotations, enabling analysis of relationships between movement, musical structure, and acoustic expression (Huang et al. 2024). PianoVAM provides synchronized piano audio, video, MIDI, and hand-pose annotations for studying finger movements and their acoustic consequences. The MMIP dataset extends this approach across instruments by combining audio, video, motion, facial expression, and MIDI information (Kyriakou et al. 2025). Additional datasets, including FürElise and VioPose, provide detailed representations of piano and violin performance gestures, respectively (Wang et al. 2024; Yoo et al. 2024).

Collective embodiment is studied through ensemble datasets, where expression emerges from interactions among performers. The Quartet Dataset provides synchronized audio, video, motion capture, and performance descriptors for string quartet recordings, supporting analysis of coordination and expressive timing (Maestre et al. n.d.). The URMP dataset provides synchronized audiovisual chamber performances with symbolic information, enabling research on instrument interaction and multimodal learning (Li et al. 2016).

Although direct embodiment datasets provide the most detailed evidence for studying gesture-related information, their limited duration, performer diversity, and instrument coverage restrict large-scale representation learning.

3.2. Indirectly Observable Embodiment Datasets

Indirectly observable embodiment datasets capture associations between music and movement without directly measuring sound-producing actions. They support analysis of audio–motion alignment and broader patterns of embodied musical expression.

Audiovisual performance datasets provide synchronized recordings of sound and visible performer movement. Although pose estimation can extract motion information from video, visible movements do not necessarily correspond to sound-producing actions. Therefore, these datasets primarily capture expressive embodiment rather than direct motor mechanisms underlying sound production.

Music-driven movement datasets provide another important category. AIST++ and related resources offer large-scale audio-synchronized human motion data for studying relationships between musical structure and expressive movement (Tsuchida et al. 2019). FineDance, AIOZ-GDANCE, and EnchantDance extend this direction to detailed choreography and multi-person interaction (Han et al. 2023; Le et al. 2023; Li et al. 2022). However, these datasets model movement generated in response to music rather than movement responsible for producing sound.

3.3. Conducting and Expressive Control Datasets

Conducting datasets represent an intermediate form of embodiment because conductor gestures influence interpretation and coordination without directly generating acoustic events. The CONDUCT dataset captures expressive conducting gestures paired with musical excerpts, enabling analysis of relationships between movement, intention, and auditory outcomes (Chen et al. 2018). The IDEA Open Movement dataset provides continuous inertial measurements of conducting behavior during performance (Lemouton et al. 2019).

These resources contribute to understanding higher-level expressive control but provide weaker evidence for inferring sound-producing gestures from audio.

3.4. Large-Scale Audio-Only Datasets

Large-scale audio datasets are the primary training resources for modern audio representation models. Although they lack explicit performer movement measurements, they provide the scale required for self-supervised and foundation-model learning.

The MAESTRO dataset provides high-resolution piano recordings aligned with MIDI, enabling analysis of expressive timing, articulation, and dynamics (Hawthorne et al. 2019). MusicNet provides structured annotations for classical music representation learning (Thickstun et al. 2017), while AudioSet and related corpora support general audio representation learning (Gemmeke et al. 2017).

These datasets may capture performance-related acoustic patterns, but without movement data, they cannot determine whether learned features represent physical gestures or correlated acoustic properties.

3.5. Physiological and Neural Datasets

Physiological and neural datasets provide complementary evidence on embodied music processing by measuring listener responses rather than performer actions. EEG datasets, including NMED-T, NMED-H, OpenMIIR, and MAD-EEG, investigate neural responses related to timing, imagery, attention, and musical structure (Cantisani et al. 2019; Kaneshiro et al. 2019; Losorelli et al. 2017; Stober et al. 2015). PMEmo and DEAP examine peripheral physiological and affective responses to music (Koelstra et al. 2012; Zhang et al. 2018).

Although these datasets do not directly address gesture-related encoding, they provide context for embodied perception and internal representations.

3.6. Dataset Taxonomy and Remaining Challenges

The relationship between dataset type and gesture-related encoding evidence is summarized in Table 1. Dataset categories differ in the type of evidence they provide: direct embodiment datasets measure physical actions, indirect embodiment datasets capture movement-related or expressive information, and audio-only datasets enable discovery

of latent gesture-related information. Physiological and neural datasets are not included in Table 1 because they characterize embodied music perception rather than performer gesture encoding.

Table 1. Representative datasets relevant to audio–gesture and embodied musical performance analysis. Embodiment indicates whether performer movement or bodily information is explicitly observed, indirectly represented, or unavailable, all website accessed on 20 August 2026.

Dataset	Embodiment	Scale	Modalities	Musical Domain	Gesture Relevance	URL
MOSA (Huang et al. 2024)	Explicit	742 performances; 23 musicians; >30 h; >570k notes	Audio + 3D MoCap + note-level semantic annotations	Piano, violin	Direct performer motion, articulation, expressive dynamics	https://zenodo.org/records/11393449
PianoVAM (Kim et al. 2025)	Explicit	106 solo recordings; 10 performers; ∼21 h	Top-view video + audio + MIDI + hand landmarks + fingering labels	Piano	Hand and finger motion; acoustic effects	https://zenodo.org/records/17811429
MMIP (Kyriakou et al. 2025)	Explicit	3.5 h	Audio + multi-angle video + 3D MoCap of body, fingers, face, and instruments + MIDI	Guitar, piano, drums	Instrument–gesture relationships and performer actions	https://zenodo.org/records/15830081
VioPose/VioDat (Yoo et al. 2024)	Explicit	Numerical scale not reported in the retrieved paper	Video + sound + calibrated 3D motion-capture poses	Violin	Bowing, hand motion, and instrument interaction	https://www.snehesh.com/viodat/
Quartet Dataset (Maestre et al. n.d.)	Explicit/ Collective	96 recordings	Separate and ambient audio tracks + video + instrumental and upper-body MoCap + bowing descriptors + score alignment	String quartet	Coordination, synchronization, and collective embodiment	https://zenodo.org/records/4774190
URMP (Li et al. 2016)	Indirect	44 arrangements: 11 duets, 12 trios, 14 quartets, 7 quintets	Individual-track audio + mixture audio/video + MIDI and PDF scores + transcription annotations	Chamber music	Isolated performer video and audiovisual synchronization; not natural ensemble interaction	https://labsites.rochester.edu/air/projects/URMP.html
AIST++ (Li et al. 2021)	Explicit	5.2 h; 1408 sequences; 1.1 M frames; 30 subjects; 10 genres	Music + 3D motion + multi-view video/ images + joints, SMPL, and camera parameters	Dance (10 genres)	Audio–motion alignment and expressive movement	https://google.github.io/aistplusplus_dataset/download.html
FineDance (Li et al. 2022)	Explicit	14.6 h; 346 paired songs and dances; 27 dancers; 22 genres	Music + FBX + SMPL/SMPL-X body and hand motion + multi-view video	Dance choreography	Music–movement correspondence and fine-grained body/hand movement	https://github.com/li-ronghui/FineDance
AIOZ-GDANCE (Le et al. 2023)	Explicit/ Collective	16.7 h; 1.808 M frames; 7 dance styles; 16 music genres	Music/audio + video frames + multi-person 3D human-mesh motion	Group dance	Collective music-driven movement	https://aioz-ai.github.io/AIOZ-GDANCE/
Choreo Spectrum3D (Han et al. 2023)	Explicit	70.32 h; 4 coarse-grained genres	Music/audio + estimated SMPL-based 3D motion + HumanML3D-format annotations	Dance	Large-scale music–movement learning from estimated pose	https://github.com/fluide1022/EnchantDance
CONDUCT (Chen et al. 2018)	Explicit/ Expressive	2 participants; 50 gesture sequences per participant; ≥5 repetitions; 1265 gesture units	Sound recordings + marker-based MoCap (34 markers, 200 Hz)	Conducting	Expressive control gestures and interpretation	https://aclanthology.org/L18-1272/

Table 1. *Cont.*

Dataset	Embodiment	Scale	Modalities	Musical Domain	Gesture Relevance	URL
MAESTRO (Hawthorne et al. 2019)	Unobserved	>172 h in the cited original release	Audio + aligned MIDI/note labels	Piano	Expressive acoustics without body data	https://magenta.withgoogle.com/datasets/maestro
AudioSet (Gemmeke et al. 2017)	Unobserved	1,789,621 labeled segments; 4971 h; 632 categories	YouTube-linked audio/video segments + human-annotated audio-event labels	General audio	Acoustic learning without gesture or body annotations	https://research.google.com/audioset/download.html

Direct instrument-performance datasets provide the strongest evidence for gesture encoding because they include synchronized measurements of performer actions, but remain limited in duration, performer diversity, and instrument coverage. For example, MOSA provides approximately 30 h of highly annotated motion-capture recordings, whereas large-scale audio datasets such as AudioSet contain millions of audio examples without direct measurements of the physical actions generating sound. Consequently, current audio representation models are typically trained at scales unavailable to motion-based datasets, while explicit embodiment information remains sparsely represented.

Current datasets also exhibit cultural and instrumental biases, with many resources focusing on Western classical instruments and expert performers. This limits the extent to which learned representations can be considered generalizable models of musical embodiment across diverse instruments, performers, and musical practices.

Instrument diversity represents an additional challenge for evaluating embodied information in audio representations. Musical instruments differ substantially in their physical mechanisms of sound production, resulting in distinct gesture–sound relationships. Bowing in string instruments, striking in percussion, and key or valve actions in keyboard and wind instruments produce different acoustic consequences. Therefore, representations learned from one instrument family may capture instrument-specific correlations rather than general principles of musical embodiment. Existing datasets are also unevenly distributed across instruments, with many focusing on piano and violin due to the availability of detailed motion measurements. Future benchmarks should therefore include greater instrument diversity and explicitly evaluate cross-instrument generalization.

Temporal scale is another limitation of current embodiment datasets. Motion-capture and instrument-performance datasets typically contain shorter recordings because synchronized movement, audio, and performance annotations require specialized equipment and controlled conditions. Although they provide detailed evidence of performer actions, their limited duration and participant diversity constrain large-scale representation learning. Conversely, large-scale audio datasets provide extensive recording diversity but only indirect evidence of the physical actions underlying sound production. This creates a fundamental trade-off: datasets with richer embodiment measurements are often smaller, whereas datasets suitable for modern representation learning provide weaker access to physical action.

Addressing this gap requires evaluation frameworks that combine large-scale pretraining with multimodal supervision. Future benchmarks should distinguish acoustic correlation from gesture-related information associated with embodied performance to determine which aspects of embodied musical processes are represented in learned models. Overall, directly observable embodiment datasets provide the strongest evidence for studying performer gesture encoding, indirectly observable datasets reveal broader sound–movement relationships, and large-scale audio corpora enable investigation of latent gesture-related information.

4. Audio Descriptors and Representations for Embodied Gesture–Music Analysis

4.1. Gesture-Relevant Audio Descriptors

Audio representations provide an indirect computational view of embodied musical gesture by capturing the acoustic consequences of performers' physical actions. Unlike motion-based representations, which directly measure bodily movement, audio representations characterize the sonic outcomes of actions such as bowing, striking, plucking, fingering, and articulation. They may therefore provide information about embodied performance processes when physical actions systematically influence timing, dynamics, articulation, pitch, or timbre.

The relevance of an audio representation to embodied gesture depends on the nature and strength of the action–sound relationship. Sound-producing gestures have relatively direct acoustic consequences: bowing affects articulation and spectral characteristics, striking influences onset and energy, and fingering contributes to pitch transitions and timing. Expressive gestures, in contrast, may influence musical interpretation through dynamics, timing, phrasing, or coordination without having a unique acoustic signature. Conducting and interpersonal gestures may have an even more indirect relationship with the resulting sound. Consequently, audio features should not be considered direct measurements of bodily movement; rather, they provide acoustic correlates whose usefulness depends on the gesture under investigation.

From an embodied perspective, the central question is therefore not simply which audio features describe a musical signal, but which acoustic properties preserve information about the physical, expressive, or coordinative processes underlying musical performance. This distinction is important because the same acoustic descriptor may provide evidence for different gesture variables depending on the musical context. For example, changes in spectral characteristics may reflect alterations in bowing or articulation during instrumental performance, whereas variations in energy may correspond to expressive intensity or physical effort. Conversely, gestures with weak or ambiguous acoustic correlates may require complementary motion, video, or physiological measurements.

Musical gesture can be considered from several complementary computational perspectives. Instrumental gestures describe physical actions involved in sound production, for which acoustic features can provide indirect measurements of technique and motor control (Traube et al. 2003). Music-induced gestures describe movements generated through listening and perception and may be influenced by acoustic properties such as loudness, pitch, and spectral characteristics (Kozak et al. 2011). Sonic gestures describe temporal sound evolution and emphasize melodic, rhythmic, and spectral changes (Dobrian 2012; Lembke 2020). These perspectives highlight that gesture-related information is not uniformly distributed in the audio signal. Rather, it emerges through temporal patterns and interactions among acoustic dimensions whose relevance depends on the embodied process being studied.

Accordingly, this section focuses on audio descriptors and learned representations according to their relevance to embodied gesture–music relationships. The discussion considers conventional acoustic descriptors, their temporal evolution, learned representations, and cross-modal approaches that explicitly relate sound to movement. Particular attention is given to the distinction between acoustic correlates of gesture and direct measurements of embodiment, since successful prediction from audio does not necessarily demonstrate that a representation has captured the underlying physical mechanism.

4.2. Relationship Between Musical Gesture and Sound

The relationship between human gesture and musical sound is fundamental to embodied musical performance because sound is not produced independently of bodily action. In instrumental performance, physical actions such as bowing, striking, fingering, and articulation directly influence the acoustic characteristics of the resulting sound. Consequently, measurable variations in timing, amplitude, pitch, and timbre can provide indirect evidence of the physical actions involved in sound production (Traube et al. 2003). This relationship provides the theoretical basis for using audio representations to investigate gesture-related information when direct movement measurements are unavailable.

The strength of the relationship between gesture and sound depends on the type of gesture being considered. Sound-producing gestures have a relatively direct relationship with acoustic outcomes because changes in physical action modify the mechanism through which sound is generated. Traube et al. demonstrated that instrumental gestures can be indirectly characterized from signal, physical, and perceptual information, illustrating how variations in playing actions can be reflected in measurable acoustic properties (Traube et al. 2003). Similarly, Bolton investigated gestural information that can be extracted from musical audio signals, supporting the use of temporal and acoustic characteristics for analysing gesture-related properties of performance (Bolton 2004).

The relationship is not limited to direct sound-producing actions. Musical sound can also influence the movements of listeners and performers through perceptual and sensorimotor processes. Kozak et al. showed that acoustic characteristics, including spectral properties, can influence gesture type and timing, demonstrating a bidirectional relationship in which sound can affect movement as well as movement influencing sound (Kozak et al. 2011). This observation is particularly relevant to music-induced and expressive gestures, for which the relationship with the acoustic signal may be indirect rather than determined by the physical mechanism of sound production.

From a computational perspective, the gesture–sound relationship can therefore be understood as a temporal mapping between movement and acoustic change. Caramiaux et al. described gesture–sound mappings in terms of instantaneous, temporal, and metaphorical relationships, emphasizing that meaningful relationships may depend on the evolution of both gesture and sound over time rather than on isolated acoustic values (Caramiaux et al. 2014). Similarly, trajectory-based approaches represent gesture and audio as time-varying multidimensional signals, allowing temporal correspondences between movement and musical events to be examined (Bevilacqua et al. 2011; Françoise et al. 2013).

Machine learning studies provide further empirical evidence that acoustic representations can contain information related to performer gestures. Best et al. demonstrated the use of audio descriptors and machine learning for musical gesture recognition (Best et al. 2018). More recent work has shown that deep learning models can recognize instrument-specific gestures from musical performance data, including violin bowing gestures, where acoustic and movement information exhibit systematic relationships (Dalmazzo and Ramirez 2019, 2021). These results indicate that gesture-related information can be inferred from musical audio, although the extent to which such information generalizes across performers, instruments, and gesture categories remains an open question.

Importantly, the existence of an acoustic correlate of a gesture does not imply that audio provides a complete representation of the underlying physical action. The same acoustic outcome may be produced by different physical strategies, while similar physical actions may result in different sounds depending on the instrument, performer, and performance context. Thus, audio should be regarded as an indirect observation of embodied musical action rather than a direct measurement of gesture. This distinction is particularly important when evaluating learned audio representations: successful prediction of a gesture-related

variable may demonstrate that acoustic information associated with that gesture is present in the representation, but does not by itself establish that the representation encodes the underlying physical mechanism.

Accordingly, the relationship between gesture and sound provides the conceptual foundation.

4.3. Audio Pre-Processing Pipeline for Gesture–Music Analysis

Audio pre-processing provides the initial computational stage for transforming musical recordings into representations suitable for gesture–music analysis. Because the relationship between performer movement and sound is inherently temporal, the pre-processing procedure should preserve the timing, spectral, and dynamic characteristics through which gesture-related information may be expressed. A general audio pre-processing workflow consists of audio standardization, segmentation and framing, time–frequency transformation, representation extraction, and normalization before the resulting representation is provided to a machine learning or deep learning model.

First, audio recordings are standardized with respect to sampling rate, channel configuration, and numerical format. Where appropriate, recordings may be converted to a common channel configuration and amplitude-normalized to reduce differences caused by recording levels. These operations should be applied carefully in gesture–music analysis because excessive normalization or signal modification may alter dynamic information that is potentially related to expressive movement and performance technique.

The standardized waveform is subsequently divided into short analysis frames, which may overlap in time. Frame-based processing allows acoustic characteristics to be examined as temporal sequences rather than as global statistics. This is particularly important for musical gesture analysis because articulation, onset characteristics, dynamics, pitch, and timbral properties evolve throughout a performance. Temporal trajectories can therefore provide a closer representation of the evolving acoustic consequences of performer actions (Bevilacqua et al. 2011; Caramiaux et al. 2014; Françoise et al. 2013).

For spectrogram-based processing, each frame can be transformed from the time domain into the frequency domain using the short-time Fourier transform (STFT). The STFT of a discrete-time audio signal $x[n]$ can be expressed as

$$X(m,k) = \sum_{n=0}^{N-1} x[n+mH]w[n]e^{-j2\pi kn/N}, \quad (1)$$

where $w[n]$ denotes the analysis window, N is the frame length, H is the hop size, m represents the time-frame index, and k represents the frequency-bin index, with $k = 0, \ldots, N-1$. The magnitude or power of $X(m,k)$ can then be used to construct a magnitude or power spectrogram that represents the evolution of frequency content over time (Miliaresi 2025).

The resulting time–frequency representation can be used directly as an input to a learning model or transformed into lower-dimensional acoustic descriptors. Common representations include magnitude or log-magnitude spectrograms, Mel spectrograms, and MFCCs. Handcrafted descriptors can additionally be extracted from the audio signal to characterize spectral, timbral, energetic, temporal, and pitch-related properties. Such descriptors have been used to establish relationships between acoustic characteristics and instrumental or expressive gestures (Best et al. 2018; Traube et al. 2003).

For waveform-based approaches, the segmented waveform can instead be supplied directly to a learning model without first constructing an explicit spectrogram. This approach retains the original temporal acoustic signal and allows the learning architecture to determine relevant transformations from the data. In contrast, spectrogram-based approaches

provide an explicit time–frequency representation before learning. Both approaches are therefore relevant to audio–gesture analysis, but they differ in the extent to which acoustic structure is predefined before representation learning.

Following representation extraction, the resulting features or representations are normalized and organized according to the requirements of the subsequent analysis. Handcrafted descriptors can be represented as feature vectors or temporal trajectories, whereas spectrograms and waveform segments can be organized as fixed-length inputs or temporal sequences for neural architectures. Temporal representations are particularly relevant to gesture–music analysis because gesture–sound relationships may depend on evolving patterns rather than isolated acoustic measurements (Caramiaux et al. 2014).

The complete processing procedure can therefore be summarized as

$$x[n] \rightarrow \tilde{x}[n] \rightarrow x_m[n]_{m=1}^{M} \rightarrow R_m \rightarrow \mathbf{z}, \tag{2}$$

where $x[n]$ is the original waveform, $\tilde{x}[n]$ is the standardized waveform, $x_m[n]$ denotes the mth analysis frame, R_m denotes the selected audio representation for the corresponding frame or segment, and $\mathbf{z}$ denotes the final representation supplied to the subsequent learning or analysis stage. Depending on the processing approach, R_m may correspond to waveform segments, spectrogram-based representations, Mel representations, MFCCs, or other acoustic descriptors.

Figure 3 summarizes the general processing workflow considered in this review. The pipeline distinguishes waveform-based, time–frequency, and descriptor-based processing while showing their common initial stages of audio standardization and temporal segmentation. The resulting representations can subsequently be used for descriptor-based analysis, machine learning models, deep learning architectures, or cross-modal audio–gesture analysis.

The audio representations discussed in this review can be described mathematically in terms of time-domain, time–frequency, spectral, energy, temporal, and pitch-related measurements. These formulations clarify how acoustic information is transformed into representations that can subsequently be used for musical gesture analysis.

The magnitude and power spectrograms introduced above are given by

$$S(m,k) = |X(m,k)|, \tag{3}$$

and

$$P(m,k) = |X(m,k)|^2. \tag{4}$$

The magnitude spectrogram represents the magnitude of the complex STFT coefficients, whereas the power spectrogram represents their squared magnitude. These representations preserve the temporal evolution of spectral content and provide the basis for several commonly used audio representations.

For perceptually motivated spectral analysis, the frequency axis can be mapped to the Mel scale. A commonly used formulation of the Mel frequency corresponding to a frequency f is

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{5}$$

Mel-frequency cepstral coefficients (MFCCs) are then obtained by applying Mel filter banks to the power spectrum, taking the logarithm of the resulting filter-bank energies,

and applying a discrete cosine transform. If E_r denotes the energy in the rth Mel filter, the cepstral coefficients can be expressed as

$$c_q = \sum_{r=1}^{R} \log(E_r) \cos\left[\frac{\pi q}{R}\left(r - \frac{1}{2}\right)\right], \tag{6}$$

where R is the number of Mel filters and q denotes the cepstral coefficient index, typically with $q = 0, \ldots, Q-1$ for the set of Q retained coefficients. MFCCs provide a compact representation of spectral characteristics and have been applied to musical gesture recognition using audio descriptors (Best et al. 2018).

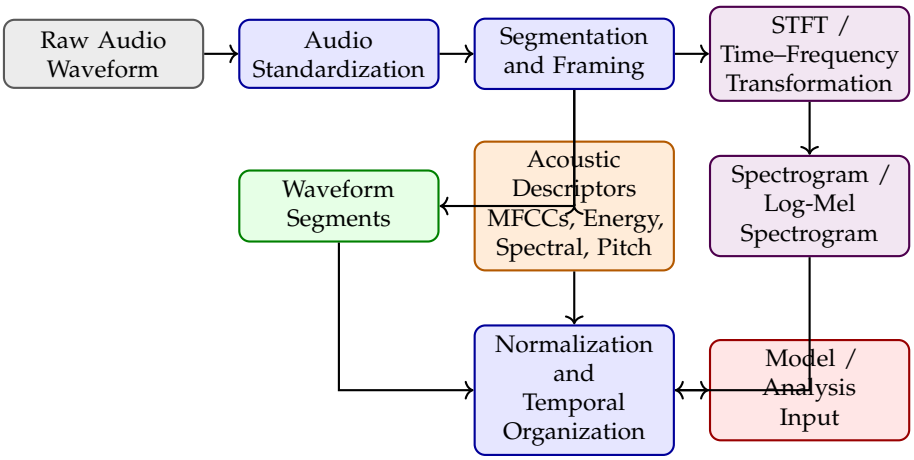


Figure 3. General audio pre-processing workflow for audio–gesture analysis. The waveform is first standardized and segmented into temporal frames. The resulting signal can be processed directly as waveform segments, transformed into time–frequency representations, or used to extract handcrafted acoustic descriptors. The resulting representations are normalized and temporally organized for subsequent analysis or machine learning and deep learning models.

Energy-related descriptors provide another representation of acoustic variation. The root-mean-square (RMS) amplitude of a frame containing N samples can be defined as

$$E_{\mathrm{RMS}}(m) = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n + mH]}. \tag{7}$$

The resulting amplitude trajectory describes changes in signal intensity over time and can therefore provide an acoustic correlate of expressive dynamics. Such temporal relationships between sound and gesture have been investigated in gesture–sound mapping studies (Caramiaux et al. 2014).

Using the magnitude spectrum, spectral centroid provides a measure of the center of mass of the spectrum and is defined as

$$C(m) = \frac{\sum_k f_k |X(m,k)|}{\sum_k |X(m,k)|}, \tag{8}$$

where f_k denotes the frequency associated with spectral bin k. Changes in spectral centroid can represent variations in spectral brightness and timbral characteristics. Such spectral properties have been investigated in relation to movement responses and gesture–sound relationships (Kozak et al. 2011; Traube et al. 2003).

Spectral flux can be used to quantify changes in spectral content between consecutive frames. It can be defined as the squared Euclidean distance between consecutive spectral frames:

$$F(m) = \sum_{k} [S(m,k) - S(m-1,k)]^2, \tag{9}$$

where $S(m,k)$ denotes the spectral representation of frame m. If a normalized spectral representation is used, the normalization should be specified as part of the feature-extraction procedure. Larger values indicate stronger spectral changes between successive frames and can therefore provide information about temporal transitions and acoustic events.

Pitch-related information can be represented through the fundamental frequency f_0. For an approximately periodic signal, the fundamental frequency is related to the fundamental period T_0 by

$$f_0 = \frac{1}{T_0}. \tag{10}$$

A sequence of estimated fundamental frequencies produces a pitch trajectory that can be analyzed in relation to melodic expression, articulation, and instrument-specific performance characteristics.

Onset-related information can similarly be represented through changes in an onset-strength function. If $O(m)$ denotes the onset-strength value at frame m, an onset event can be associated with a local increase in $O(m)$. The resulting onset trajectory provides information about the temporal organization of musical events and can be used to characterize articulation and timing.

These mathematical representations illustrate that audio-based gesture analysis does not depend on a single acoustic descriptor. Rather, gesture-related information may emerge from temporal relationships among several acoustic dimensions, including energy, spectral content, pitch, and event timing. This is consistent with the view that gesture–sound relationships are multidimensional and evolve over time (Caramiaux et al. 2014; Kozak et al. 2011; Traube et al. 2003). Consequently, audio representations used for gesture analysis may consist of individual descriptors, multidimensional descriptor trajectories, spectrograms, or learned representations derived from these acoustic inputs.

4.4. Acoustic Correlates of Embodied Musical Gesture

The relationship between embodied gesture and audio can be understood through the acoustic consequences of physical actions. During instrumental performance, bodily movements are translated into sound through instrument-specific mechanical processes. Bow pressure and velocity, for example, influence the spectral and dynamic characteristics of bowed-string sounds, while striking actions affect onset characteristics, amplitude, and spectral energy. Similarly, fingering actions influence pitch transitions, articulation, and timing. These relationships provide the basis for using audio descriptors as indirect indicators of embodied performance variables (Bolton 2004; Traube et al. 2003).

The strength of this correspondence varies considerably across gesture types. Sound-producing actions generally exhibit stronger acoustic coupling because the physical action directly contributes to sound generation. Expressive gestures may instead influence the interpretation of a musical passage through changes in timing, dynamics, articulation, or phrasing. In such cases, acoustic descriptors may capture consequences of the gesture without uniquely identifying the physical movement that produced them. This distinction is particularly important when interpreting audio-based gesture recognition results, because an accurate prediction of a gesture-related variable does not necessarily imply that the underlying bodily movement has been explicitly represented.

The acoustic signal can therefore be regarded as an observation of embodied musical activity rather than a direct representation of the body. This perspective motivates the selection of descriptors according to the specific gesture–sound relationship under investigation. Spectral descriptors may be informative for gestures that modify timbre and articulation, energy descriptors for actions associated with dynamics or expressive intensity, temporal descriptors for timing and articulation, and pitch-related descriptors for actions affecting pitch production and melodic shaping.

Prior work has employed handcrafted audio descriptors to characterize acoustic dimensions that may preserve information about embodied performance. These descriptors remain relevant because they provide interpretable relationships between measurable sound properties and aspects of musical action. Rather than treating them as generic audio features, their value for gesture analysis lies in the extent to which changes in these properties correspond to changes in physical action, expressive behavior, or coordination (Table 2).

Table 2. Audio descriptor categories and their relevance to embodied gesture–music analysis.

Audio Categories	Representative Descriptors	Embodied Gesture Relevance
Spectral and timbral descriptors	MFCCs, spectral centroid, spectral flux, spectral envelope	Acoustic consequences of bowing, striking, articulation, contact, playing technique, and timbral variation
Energy and dynamic descriptors	RMS energy, loudness, amplitude envelope, dynamic range	Expressive intensity, dynamic shaping, attack force, and acoustic correlates of physical effort
Temporal and rhythmic descriptors	Onset strength, attack time, duration, tempo, rhythmic descriptors	Timing of sound-producing actions, articulation, beat alignment, and coordination patterns
Pitch and harmonic descriptors	Fundamental frequency, pitch contour, harmonic descriptors	Fingering, pitch control, melodic shaping, and pitch-related motor actions
Temporal feature trajectories	Time-varying acoustic feature sequences	Evolution of gesture phases, transitions, articulation, and expressive shaping
Learned audio representations	CNN embeddings, recurrent/Transformer features, self-supervised embeddings	Distributed representations potentially encoding performer-, action-, and performance-related information

4.4.1. Spectral and Timbral Descriptors

Spectral and timbral descriptors characterize acoustic properties that can be influenced by instrument mechanics, articulation, and playing technique. The Timbre Toolbox provides a framework for extracting spectral, spectro-temporal, harmonic, and energy-related features, emphasizing the multidimensional nature of timbre (Peeters et al. 2011). For embodied gesture analysis, these descriptors are relevant when physical actions modify the spectral distribution or temporal evolution of sound.

Descriptors such as spectral centroid, spectral flux, fundamental frequency, and spectral envelope capture brightness, spectral variation, pitch evolution, and harmonic structure. Variations in these descriptors can provide indirect evidence of changes in playing technique. Traube et al. demonstrated that changes in instrumental technique produce measurable spectral variations, establishing a link between physical actions and their acoustic consequences (Traube et al. 2003). MFCCs provide compact representations of spectral structure and have also been employed in musical gesture recognition and sound classification (Best et al. 2018).

The interpretation of spectral features should nevertheless remain context-dependent. A change in spectral centroid, for example, does not correspond to a unique physical movement. It may result from changes in bow position, articulation, instrument resonance, or recording conditions. Consequently, spectral descriptors are most informative when interpreted in relation to the gesture, instrument, and performance context.

4.4.2. Energy and Dynamic Descriptors

Energy-related descriptors provide information about variations in acoustic intensity that may correspond to expressive movement and sound-producing actions. Short-time energy, RMS energy, loudness, amplitude envelopes, and dynamic range characterize changes in intensity over time and can therefore provide acoustic correlates of movement amplitude, physical effort, attack force, and expressive shaping.

Caramiaux et al. demonstrated that energy trajectories can support perceptually meaningful gesture–sound mappings, illustrating how changes in sound intensity can be related to changes in gesture variables (Caramiaux et al. 2014). These relationships are particularly relevant for expressive performance, where changes in physical action may be reflected in dynamic shaping rather than in discrete musical events.

However, energy should not be interpreted as a direct measure of physical effort. Acoustic intensity depends on instrument characteristics, microphone position, room acoustics, and performance technique. Thus, energy-based descriptors are better understood as acoustic correlates whose relationship with bodily movement must be established for the specific performance context.

4.4.3. Temporal and Rhythmic Descriptors

Temporal and rhythmic descriptors are particularly important for embodied gesture analysis because musical gestures unfold over time and are often coordinated with rhythmic structures. Onset strength, attack time, note duration, tempo, and rhythmic descriptors characterize the temporal organization of musical events and can provide information about the timing of sound-producing actions and expressive coordination.

Onset-related features can reveal gesture-related characteristics associated with instrumental technique and articulation. Bolton demonstrated that onset characteristics can be used to identify differences associated with instrumental gestures and techniques (Bolton 2004). More generally, temporal representations are important for examining how movements and musical events are synchronized because gesture-related information is typically expressed through evolving temporal patterns rather than isolated acoustic values.

Temporal descriptors are also relevant to cross-modal mappings between movement and sound. Caramiaux et al. distinguished instantaneous and temporal gesture–sound mappings, emphasizing that meaningful relationships may depend on the evolution of both modalities over time (Caramiaux et al. 2014). This temporal perspective is consistent with studies showing that movement responses to music can depend on temporal and spectral characteristics of the stimulus (Kozak et al. 2011).

4.4.4. Pitch and Harmonic Descriptors

Pitch and harmonic descriptors characterize melodic and harmonic properties that may be related to embodied actions such as fingering, pitch control, and melodic shaping. Fundamental frequency and pitch contours provide time-varying representations of pitch, while harmonic descriptors characterize the distribution and evolution of spectral components.

For sound-producing gestures, pitch-related descriptors can provide indirect evidence of motor actions when physical changes in fingering or instrument control produce measurable pitch transitions. However, as with spectral and energy descriptors, the mapping is not one-to-one. The same pitch trajectory may be produced by different combinations of physical actions, while similar actions may produce different acoustic outcomes across instruments and performers. Pitch descriptors are therefore most informative when combined with temporal, spectral, or motion-based information.

4.5. Temporal Representations of Gesture–Music Relationships

Because embodied musical gestures are dynamic processes, audio descriptors are generally more informative when represented as temporal trajectories rather than as isolated values or global statistics. Musical sounds evolve through onset, development, transition, and decay, while physical gestures similarly consist of preparation, action, and recovery phases. Preserving these temporal structures is therefore essential for modeling relationships between sound and movement.

Caramiaux et al. distinguished instantaneous, temporal, and metaphorical gesture–sound mappings (Caramiaux et al. 2014). In temporal mappings, evolving structures in the gesture and audio streams are compared using representations such as loudness envelopes, spectral trajectories, and pitch contours. Similar trajectory-based approaches have been used to represent movement as multidimensional sequences and to preserve temporal relationships between gesture and sound (Bevilacqua et al. 2011; Françoise et al. 2013).

This temporal perspective also helps distinguish gesture-related acoustic information from generic musical descriptors. A single value such as average loudness may characterize a recording but provide limited information about when and how a gesture occurred. In contrast, the temporal evolution of loudness, spectral energy, pitch, or onset characteristics can reveal transitions and coordination patterns that are more directly relevant to embodied performance.

Consequently, gesture–music analysis should consider not only which acoustic dimensions are represented but also how their temporal organization corresponds to gesture phases and musical events. This principle provides a bridge between conventional descriptor-based approaches and sequential or learned representations.

4.6. Learned Audio Representations for Embodied Gesture Analysis

Machine learning approaches provide mechanisms for transforming acoustic signals into representations suitable for gesture recognition, prediction, retrieval, and cross-modal modeling. Early approaches relied primarily on multidimensional handcrafted feature spaces, where descriptors such as spectral centroid, spectral flux, loudness, and MFCCs were used to characterize musical events and recognize gesture-related categories (Best et al. 2018). Zbyszynski et al. demonstrated that timbral descriptors can also act as intermediate representations linking physical actions and sonic outcomes through machine-learning-based gesture–timbre mappings (Zbyszyński et al. 2019).

Deep learning extends this approach by learning hierarchical representations from spectrograms, time–frequency representations, or raw waveforms. CNNs, recurrent networks, and Transformer-based architectures can capture local acoustic patterns as well as longer temporal and contextual dependencies (Zhu et al. 2023). For embodied gesture analysis, their principal advantage is the ability to model nonlinear and temporally distributed relationships between acoustic signals and performer actions without requiring every relevant tasks. Dalmazzo et al. used CNN and CNN–LSTM model improvements from deep architectures, which should not be interpreted as evidence that learned audio representations explicitly encode bodily mechanisms. A model may successfully infer a gesture-related category by exploiting acoustic regularities associated with performer identity, instrument characteristics, or a descriptor to be manually specified.

Several studies demonstrate the potential of learned audio representations for gesture-related tasks. Dalmazzo et al. used CNN and CNN–LSTM models for violin bow-stroke recognition using synchronized audio and motion data (Dalmazzo and Ramirez 2021). Kritsis et al. applied CNN–LSTM architectures to virtual musical instrument gesture recognition and reported high recognition performance, while also observing reduced performance across performers, highlighting the challenge of generalization (Kritsis et al.

2018). Other studies similarly report improvements from deep architectures in modeling relationships between acoustic and movement patterns (Bacher 2020; Dalmazzo and Ramirez 2019; Rhodes et al. 2020).

These results, however, should not be interpreted as evidence that learned audio representations explicitly encode bodily mechanisms. A model may successfully infer a gesture-related category by exploiting acoustic regularities associated with performer identity, instrument characteristics, articulation, or musical style. The distinction between predicting a gesture-related variable and representing the underlying embodied process is therefore important.

Self-supervised and representation-learning approaches provide an additional mechanism for extracting latent structures from audio without requiring gesture labels. Models such as wav2vec 2.0 and HuBERT learn general-purpose acoustic representations from large-scale unlabeled data through predictive and contrastive objectives (Baevski et al. 2020; Hsu et al. 2021). These representations capture information related to temporal organization, timbre, pitch, and other acoustic properties, although such information is distributed across latent dimensions rather than explicitly represented by interpretable features (Cramer et al. 2021).

For embodied gesture analysis, the important question is whether these latent acoustic structures also preserve information associated with performer actions. Because self-supervised audio models are generally optimized for acoustic objectives rather than physical-action modeling, embodiment-related information cannot be assumed from their performance on conventional audio tasks. Instead, it must be evaluated through dedicated probing, prediction, or cross-modal experiments.

Perceptually informed representation-learning approaches provide another way of investigating the relationship between latent representations and meaningful acoustic dimensions. Metric-learning and perceptual constraints can improve the correspondence between embedding distances and perceptual similarity, including timbral and acoustic attributes. However, these properties are typically distributed across multiple latent variables rather than encoded in individually interpretable dimensions.

For embodied music analysis, this lack of interpretability creates an important methodological challenge. An embedding may contain information sufficient to predict a performer's action while providing little insight into which acoustic characteristics carry that information. Consequently, latent representations should be evaluated not only according to predictive performance but also according to their robustness across performers, instruments, recording conditions, and musical contexts.

Cross-Modal Representation Learning

Cross-modal learning provides a particularly relevant framework for embodied audio representation because it explicitly relates acoustic signals to complementary measurements of physical action. Such approaches can align audio with motion capture, video, symbolic performance data, or physiological measurements and thereby provide additional supervision for learning gesture-related representations (Morgado et al. 2021; Yu et al. 2022).

The motivation for cross-modal learning follows directly from the indirect nature of audio-based embodiment. Audio datasets can be collected at substantially larger scales than synchronized motion datasets, but audio contains physical information only through its acoustic consequences. Conversely, motion recordings provide detailed measurements of bodily movement but are generally more expensive to acquire and annotate. Learning from synchronized audio–movement pairs can therefore combine the scalability of audio with the explicit embodiment information provided by motion measurements.

Cross-modal objectives may encourage representations to retain information shared between sound and movement. Contrastive learning, representation alignment, and distillation approaches can, for example, encourage an audio embedding to remain informative about gesture variables that are also observable in motion data. Such approaches are particularly relevant for expressive timing, articulation, action-related variation, and coordination between performer movement and musical events.

Nevertheless, cross-modal learning does not automatically solve the ambiguity of gesture–sound relationships. Its effectiveness depends on the quality and temporal alignment of the paired modalities and on whether the recorded movement represents a meaningful component of musical embodiment. Evaluation should therefore consider generalization across performers, instruments, and performance contexts rather than relying solely on reconstruction, classification, or retrieval accuracy.

4.7. Limitations of Audio-Based Embodiment

The use of audio representations for embodied gesture analysis is fundamentally constrained by the non-unique relationship between movement and sound. Different gestures can produce similar acoustic outcomes, while similar gestures can generate different sounds depending on instrument mechanics, performer characteristics, and performance context (Leman 2007; Maes 2016). Consequently, the presence of gesture-related information within an audio representation does not necessarily indicate that the representation captures the physical mechanisms underlying sound production.

This distinction is particularly important when comparing different types of musical gesture. Sound-producing actions, such as articulation, bowing, striking, and plucking, generally leave stronger acoustic traces because they directly influence timing, dynamics, and timbral properties. Expressive or communicative gestures, including posture, body movement, and visual interaction between performers, often have weaker or less consistent acoustic correlates. Ensemble performance introduces further ambiguity because the audio signal reflects collective musical activity while obscuring the contribution of individual performers and their underlying motor actions (Papiotis et al. 2014).

Audio representations are also sensitive to factors that are unrelated to embodiment. Performer identity, instrument type, recording conditions, microphone placement, room acoustics, and musical style can all influence the acoustic signal. A model may consequently learn correlations that support gesture prediction without learning a representation that generalizes across these factors. This issue is particularly relevant for deep and self-supervised representations, where latent features may encode multiple sources of variation simultaneously.

Therefore, demonstrating embodied information in an audio representation requires more than achieving high performance on a gesture-related prediction task. Evaluation should investigate whether the representation remains informative when performer identity, instrument, recording conditions, and musical material vary. Where possible, audio-based analysis should also be complemented by motion or other embodied measurements to distinguish genuine gesture–sound relationships from indirect statistical correlations.

Overall, audio representations are best understood as complementary rather than equivalent alternatives to direct movement measurements. Their principal value lies in capturing acoustic consequences of embodied musical actions and in providing scalable representations that can be related to movement through temporal and cross-modal modeling. This perspective places audio descriptors within a broader embodied analysis framework, in which the correspondence between physical action, sonic outcome, and perceptual or expressive function is treated as an explicit object of study.

5. Deep Learning for Embodied Audio–Gesture Analysis

Deep learning approaches provide a framework for modelling relationships between musical sound and performer actions, particularly when the acoustic consequences of physical gestures are difficult to characterize using predefined descriptors. From an embodied perspective, their relevance lies not simply in their ability to classify musical sounds or gestures, but in their capacity to learn representations that preserve information about the relationship between performer action, instrument interaction, and resulting sound. Unlike traditional machine learning methods based on handcrafted features, deep architectures learn hierarchical representations directly from acoustic or multimodal inputs, enabling the modelling of complex spatiotemporal relationships between sound and gesture-related information. Recent studies report recognition accuracies exceeding 90% for instrument-specific gesture recognition tasks, with CNN, recurrent, and hybrid architectures consistently outperforming conventional classifiers such as Hidden Markov Models (HMMs), support vector machines (SVMs), and k-nearest neighbours (k-NN) (Bacher 2020; Dalmazzo and Ramirez 2021; Kritsis et al. 2018; Rhodes et al. 2020).

However, high recognition performance should not by itself be interpreted as evidence that a model has learned an embodied representation. Although deep learning approaches achieve strong recognition performance in musical gesture tasks, direct comparison between reported accuracies remains challenging. Existing studies use heterogeneous datasets that differ in instrument type, gesture definition, recording conditions, performer number, modality availability, and evaluation protocols. Consequently, recognition rates cannot be interpreted as a unified measure of progress in embodied audio representation learning. A model achieving high accuracy for a specific instrument or performance context may exploit instrument-dependent acoustic cues rather than learning generalizable gesture-related representations. Furthermore, the limited availability of large-scale audio–gesture datasets remains a fundamental bottleneck, restricting cross-instrument evaluation and systematic assessment of representation generalization.

Beyond architectural choices, the effectiveness of deep learning approaches depends strongly on the learning objective used to shape the learned representation. Supervised approaches optimize task-specific objectives, such as gesture classification or performance parameter prediction, and therefore learn representations directly associated with annotated variables. In contrast, self-supervised approaches learn general-purpose audio representations through objectives such as masked prediction, contrastive learning, or predictive coding, allowing models to discover acoustic structures from large-scale unlabeled data.

For embodied musical gesture analysis, however, the critical question is whether the learned acoustic structure contains information related to performer action rather than only information useful for describing the sound itself. Cross-modal objectives further extend this framework by aligning audio representations with complementary modalities, such as motion, video, or symbolic performance information, encouraging models to capture relationships between sound and embodied action. Consequently, evaluating embodied information in learned representations requires consideration not only of the neural architecture, but also of the objective that determines which aspects of musical sound are emphasized during learning.

5.1. Deep Learning for Sound-Producing Gestures

The strongest evidence for audio-based modelling of embodied musical gestures currently comes from sound-producing actions, for which physical movements have relatively direct acoustic consequences. Such actions include bowing, striking, plucking, fingering, and articulation. In these cases, performer movement affects timing, spectral characteristics,

dynamics, and other measurable properties of the resulting sound, providing an acoustic trace that can be exploited by machine learning models.

The strongest evidence currently comes from instrument-specific studies, particularly violin performance analysis. Dalmazzo et al. achieved accuracies of 97.15% using CNNs, 98.55% using 3D Multi-Head CNNs, and 99.23% using CNN–LSTM models for bow-stroke recognition from synchronized audio and motion recordings (Dalmazzo and Ramirez 2021). These results represent some of the highest reported performances in musical gesture recognition; however, they should be interpreted within the specific context of violin bowing, where the relationship between physical action and acoustic outcome is relatively well structured and strongly reflected in timbral and temporal variations. This example illustrates the potential of learned acoustic representations to discriminate between different sound-producing actions, while also demonstrating the importance of synchronized movement information for establishing the relationship between action and sound. Similar performance cannot be assumed for other instruments or gesture categories, particularly expressive body movements and collective performance interactions, where gesture–sound mappings are more ambiguous.

Related studies have demonstrated the usefulness of learned acoustic representations for modelling other instrumental and virtual-instrument gestures. Kritsis et al. applied CNN–LSTM architectures for virtual musical instrument gesture recognition, achieving high recognition rates while observing reduced performance across performers, highlighting challenges in generalization (Kritsis et al. 2018). Comparative studies further indicate that deep architectures can outperform conventional classifiers by learning nonlinear relationships between acoustic and movement patterns (Bacher 2020; Dalmazzo and Ramirez 2019; Rhodes et al. 2020). These findings suggest that deep models can capture acoustic regularities associated with performance actions, but they also emphasize that performer-independent and instrument-independent evaluation is necessary before such representations can be considered transferable models of embodied musical action.

5.2. Multimodal Learning and Embodied Information

Moreover, many high-performing systems rely on multimodal information, including motion capture, inertial sensors, video, or wearable devices. These modalities provide direct access to physical movement, allowing models to exploit information unavailable from audio alone. Therefore, high recognition accuracy in multimodal settings does not demonstrate that acoustic representations independently contain embodied gesture information.

Multimodal learning is nevertheless particularly valuable for embodied musical gesture analysis because it provides an explicit correspondence between the acoustic outcome and the physical action that generated or accompanied it. For example, synchronized audio and motion recordings can allow a model to learn which acoustic changes correspond to changes in bowing, striking, articulation, or other performance actions. Cross-modal objectives can therefore be used not only to improve recognition, but also to encourage the learned audio representation to retain information that is shared with movement. This is particularly relevant when investigating latent audio embeddings, where gesture-related information may be distributed across multiple dimensions rather than represented by an individual interpretable feature.

At the same time, multimodal performance should be distinguished from audio-only representation learning. If movement, video, or sensor information is directly available to the model, high gesture-recognition accuracy may result from explicit access to physical information rather than from the acoustic representation. Multimodal results therefore provide evidence for the existence of an audio–gesture relationship, but do not

necessarily demonstrate that the audio representation alone encodes the corresponding embodied process.

5.3. Learned Audio Representations and Generalization

Recent self-supervised models, including wav2vec 2.0 and HuBERT, learn general-purpose acoustic representations from large-scale unlabeled data through predictive and contrastive objectives (Baevski et al. 2020; Hsu et al. 2021). These models capture information related to rhythm, timbre, pitch, and temporal organization, although such properties are distributed across complex latent spaces rather than represented by individual interpretable dimensions (Cramer et al. 2021).

For embodied musical gesture analysis, the potential value of such representations lies in their ability to retain acoustic information that indirectly reflects performer action. However, because self-supervised objectives are generally optimized for acoustic prediction rather than physical performance modelling, the presence of gesture-related information cannot be assumed. A learned embedding may encode articulation, timing, or timbral characteristics that correlate with a particular gesture while providing little information about the underlying physical mechanism.

The choice of acoustic input representation remains important for model performance. Mel spectrograms are widely used because their time–frequency structure is suitable for CNN-based learning, while MFCCs remain valuable for timbral modelling and raw waveform approaches are increasingly adopted in self-supervised learning systems (Humphrey et al. 2012; Moysis et al. 2023; Pons et al. 2016). Thus, current approaches represent a transition from manually designed acoustic descriptors toward learned representations, while retaining classical features when interpretability is required.

An important consideration is therefore whether learned representations generalize across the factors that influence the acoustic realization of a gesture. Despite their strong performance, models may exploit correlations associated with performer identity, instrument characteristics, recording conditions, or musical style. For example, a model trained to recognize bowing gestures from one violinist may learn characteristics specific to that performer or instrument rather than properties of bowing that generalize across performers. Evaluation across performers, instruments, recording environments, and musical contexts is consequently essential for determining whether learned representations encode transferable gesture-related information.

5.4. Latent Audio Representations and Embodied Information

Learned audio embeddings provide a potential mechanism for capturing gesture-related information because they represent distributed acoustic structures emerging from the relationship between sound production and performance behavior. Unlike handcrafted descriptors, these representations do not explicitly define acoustic dimensions but may preserve latent relationships among timing, articulation, dynamics, timbre, and the physical actions associated with sound production.

Representation-learning studies indicate that learned embeddings preserve meaningful acoustic and perceptual structures. Self-supervised models trained through predictive and contrastive objectives can encode fundamental frequency, spectral characteristics, and temporal organization without explicit supervision. Analyses of wav2vec-based representations show that different network layers capture different levels of acoustic information, with earlier layers representing local signal properties and deeper layers encoding more abstract temporal and contextual structures.

For embodied gesture analysis, this layered organization raises an important question: at which representational level, if any, is information about performer action retained?

Lower-level representations may preserve acoustic correlates of articulation and sound production, whereas higher-level representations may emphasize musical context and temporal organization. Determining this relationship requires targeted probing rather than assuming that deeper or more task-specific representations necessarily contain more embodied information.

Although self-supervised audio models learn powerful representations from large-scale data, their optimization objectives target acoustic prediction rather than explicit modeling of physical performance processes. Therefore, whether embodiment-related information naturally emerges from audio-only learning remains an empirical question requiring dedicated evaluation.

Perceptually informed representation-learning approaches further investigate whether latent spaces correspond to human auditory judgments. Metric-learning and perceptual constraints can improve alignment between embedding distances and perceptual similarity measures, including timbral and acoustic attributes. However, acoustic properties are rarely represented along individual interpretable dimensions; instead, they are distributed across multiple latent variables and network components. Consequently, the relationship between embedding structure and perceptual or physical meaning remains complex.

5.5. Cross-Modal Learning for Embodied Audio Representations

Cross-modal learning provides a promising framework for incorporating embodiment-related information into audio representation learning by aligning acoustic signals with complementary modalities, including motion capture, video, symbolic performance data, and physiological measurements (Morgado et al. 2021; Yu et al. 2022). Unlike audio-only approaches, multimodal objectives explicitly model relationships between sound and physical action, providing additional supervision for learning gesture-related representations.

For musical applications, cross-modal learning addresses the fundamental scale–fidelity trade-off in embodied datasets. Motion-based recordings provide detailed measurements of performer movement but remain limited in scale and diversity, whereas large-scale audio datasets provide extensive coverage but contain embodiment only indirectly. By learning shared representations from aligned audio and movement data, cross-modal approaches can encourage audio embeddings to retain gesture-related information through contrastive learning, representation distillation, or shared embedding objectives.

Such approaches may enable audio representations to capture performance characteristics beyond traditional acoustic descriptors, including expressive timing, articulation patterns, and action-related variations. However, their effectiveness depends on dataset quality, temporal alignment, and whether recorded movements represent meaningful aspects of musical embodiment. Future evaluation should therefore consider not only reconstruction or retrieval performance, but also whether learned audio embeddings generalize across performers, instruments, and performance contexts.

5.6. Limitations of Learned Audio Representations

Despite their ability to capture complex acoustic structures, learned audio representations are constrained by the non-unique relationship between movement and sound. Different gestures may produce similar acoustic outcomes, while similar gestures may generate different sounds depending on instrument mechanics, performer characteristics, and performance context (Leman 2007; Maes 2016). Therefore, the presence of gesture-related information within an embedding does not necessarily indicate that the model has learned embodied mechanisms or the physical processes underlying sound production.

This distinction is fundamental for musical gesture analysis. Sound-producing gestures, such as articulation, bowing, or striking, generally leave stronger acoustic traces

because they directly influence timing, dynamics, and timbral properties. In contrast, expressive or communicative gestures, including posture, body movement, or visual interaction between performers, often have weaker or inconsistent acoustic correlates. Ensemble performance introduces additional ambiguity because audio reflects collective coordination while obscuring contributions from individual performers and their underlying motor actions (Papiotis et al. 2014).

Consequently, learned representations may infer gesture-related variables by exploiting statistical relationships between acoustic features and performance characteristics without explicitly representing the underlying embodied processes. For example, an embedding may capture bowing style, articulation, or expressive timing because these factors influence the resulting sound, while lacking explicit information about the physical mechanics of bow movement. Demonstrating embodied information therefore requires evaluation methods that test whether learned representations capture meaningful gesture variables beyond instrument-, performer-, or recording-specific acoustic correlations.

5.7. Explainability of Deep Learning Approaches

Despite their strong predictive performance, deep learning models provide limited insight into which aspects of acoustic information correspond to embodied musical gesture. This lack of interpretability represents a significant challenge because high recognition accuracy does not necessarily indicate that models have learned meaningful relationships between sound and physical action. Neural networks may instead exploit spurious correlations associated with instrument identity, performer characteristics, recording conditions, or other dataset-specific factors. Therefore, explainability methods are increasingly important for investigating whether learned audio representations capture information related to physical performance processes.

Several approaches have been proposed to analyze the reliability and meaning of learned representations. Kim et al. (Kim et al. 2019) investigated the trustworthiness of deep music embeddings by examining whether semantically related musical examples maintain meaningful relationships within learned latent spaces. For embodied gesture research, representation-based analyses can help determine whether performances involving similar physical actions exhibit consistent organization within an acoustic embedding space. If an embedding contains gesture-related information, acoustically different realizations of similar physical actions should exhibit meaningful organization rather than arbitrary proximity.

Another important role of explainability is identifying whether deep models rely on meaningful acoustic cues or unintended dataset biases. Haunschmid et al. (Haunschmid et al. 2020) introduced audioLIME, an explanation approach for audio models that relates predictions to interpretable audio components rather than only abstract neural activations. This is particularly relevant for musical gesture analysis, where a model may achieve high accuracy by exploiting instrument-specific timbral characteristics instead of learning acoustic consequences associated with physical action. Similarly, research on "Clever Hans" predictors has demonstrated that deep networks can achieve strong predictive performance by relying on spurious features that are unrelated to the intended task (Anders et al. 2019; Lapuschkin et al. 2019). These findings highlight the importance of verifying that gesture recognition models learn meaningful action-related information rather than accidental correlations present in training datasets.

A central question for embodied audio analysis is therefore not only which acoustic cues influence model predictions, but whether these cues correspond to plausible physical mechanisms linking performer action and sound production.

Explainability can also provide evidence that deep models capture acoustically meaningful aspects associated with sound production. Wang et al. (Wang et al. 2023) applied layer-wise relevance propagation to audio classification of instrumental playing techniques and showed that model decisions were associated with acoustic properties related to the physical characteristics of performance techniques. Such analyses are especially relevant for embodied musical gesture research because they establish connections between learned acoustic features and aspects of sound production, including technique-dependent variations in articulation and timbre.

However, explainability methods alone cannot demonstrate that a model has learned a representation that reflects embodied processes. Identifying important acoustic regions or influential features does not necessarily prove that these features encode physical gestures, as they may still reflect instrument identity, performer characteristics, or recording conditions. Consequently, explainability should be considered a complementary approach for analysing deep audio representations and should be combined with cross-modal comparison, perturbation-based testing, and cross-domain evaluation. These combined strategies are necessary to determine whether learned representations capture transferable relationships between musical sound and human action rather than dataset-specific acoustic regularities.

5.8. Case Studies in Embodied Musical Gesture Analysis

The relationship between musical sound and embodied gesture has been investigated through a range of complementary computational approaches. These studies illustrate that musical gesture analysis is not limited to gesture recognition, but also includes the extraction of gesture-related information from audio, the mapping between movement and sound, the analysis of instrumental technique, and the learning of representations that relate acoustic and physical performance characteristics. Table 3 summarizes representative examples and highlights the different ways in which embodiment has been operationalized computationally.

An early example is provided by Traube et al., who investigated the indirect acquisition of instrumental gesture from signal, physical, and perceptual information (Traube et al. 2003). Their work illustrates a fundamental principle of audio-based gesture analysis: physical actions do not need to be measured directly if their effects can be identified in the resulting acoustic signal. This perspective is particularly relevant to sound-producing gestures, for which changes in articulation, spectral characteristics, and other acoustic properties can provide indirect evidence of instrumental action.

Kozak et al. examined the relationship between spectral characteristics of sound and the timing and type of movement produced in response to musical stimuli (Kozak et al. 2011). Their study demonstrates the complementary relationship between acoustic and movement representations, showing that perceptually salient sound characteristics can influence the temporal and qualitative properties of movement. This provides an example of gesture analysis from a perceptual rather than purely production-oriented perspective.

Caramiaux et al. investigated mappings between gesture and sound through listening and distinguished between instantaneous, temporal, and higher-level relationships between the two modalities (Caramiaux et al. 2014). Their work demonstrates that gesture–sound relationships cannot always be represented through one-to-one mappings between individual features. Instead, meaningful relationships may emerge from the temporal evolution of both movement and sound. Similarly, Françoise et al. explored gesture–sound mapping by demonstration, showing how examples of movement and sound can be used to learn mappings for interactive musical systems (Françoise et al. 2013). These studies

are particularly relevant to embodied analysis because they treat movement and sound as temporally structured, interacting signals rather than independent data streams.

Table 3. Representative case studies of musical gesture analysis using audio and/or multimodal information.

Study	Instrument/ Context	Gesture Focus	Audio/ Multimodal Data	Main Contribution
Traube et al. (2003)	Piano	Expressive gestures and body movement	Audio and video	Analysis of indirect relationships between expressive movement and acoustic features.
Kozak et al. (2011)	Piano	Upper-body and performance gestures	Audio and motion	Investigation of the relationship between performer movement and musical expression.
Caramiaux et al. (2014)	Musical performance	Gestural control	Audio and gesture	Mapping between performer gestures and sound-related parameters.
Bolton (2004)	Musical performance	Expressive gesture	Audio and movement	Study of gesture as a component of musical communication and expression.
Bevilacqua et al. (2011)	Musical performance	Gesture–sound relationships	Audio and motion	Analysis of gestural information for musical interaction and expressive control.
Françoise et al. (2013)	Interactive music	Musical gestures	Audio and motion	Frameworks for learning mappings between movement and sound.
Best et al. (2018)	Musical performance	Expressive movement	Audio and video	Investigation of relationships between performer movement and musical characteristics.
Zbyszyński et al. (2019)	Instrumental performance	Instrumental and expressive gestures	Audio and motion capture	Analysis of gesture in relation to musical performance and interaction.
Dalmazzo and Ramirez (2019)	Violin	Bowing gesture	Audio and motion	Study of bowing movements and their acoustic consequences.
Dalmazzo and Ramirez (2021)	Violin	Bowing and performance gesture	Audio and motion	Application of multimodal analysis to relate movement characteristics to sound.
Kritsis et al. (2018)	Musical performance	Performance gestures	Audio and motion	Investigation of gesture-aware musical interaction using multimodal information.
Rhodes et al. (2020)	Musical performance	Expressive and instrumental gestures	Audio and movement	Analysis of performer movement in relation to musical sound and expression.

Instrument-specific studies provide more direct examples of computational gesture recognition. Dalmazzo and Ramírez investigated bowing gestures in violin performance using machine learning methods (Dalmazzo and Ramirez 2019). Their work focused on distinguishing different bowing-related movement patterns and demonstrates how machine learning can be used to model the relationship between physical performance technique and musical output. Dalmazzo et al. subsequently extended this line of research using deep learning to estimate patterns of musical gesture from synchronized audio and motion recordings (Dalmazzo and Ramirez 2021). CNN, 3D CNN, and CNN–LSTM architectures were evaluated for violin bow-stroke recognition, with the results demonstrating the ability of deep models to learn complex relationships between acoustic and movement information.

Best et al. investigated musical gesture recognition using machine learning methods and audio descriptors (Best et al. 2018). This study represents an audio-centered approach in which handcrafted acoustic descriptors are used as computational representations for gesture recognition. It therefore provides a useful contrast with multimodal approaches based on synchronized motion data and illustrates the potential of audio to provide indirect information about musical actions.

Other studies have focused on gesture–sound interaction rather than recognition alone. Zbyszynski et al. developed a gesture–timbre space in which multidimensional gesture features were mapped to timbral characteristics using machine learning and concatenative synthesis (Zbyszyński et al. 2019). This work illustrates how embodied movement can be treated as a multidimensional control space and related to perceptually meaningful properties of sound. Rather than asking whether a particular gesture can be recognized, the approach investigates how continuous movement dimensions can be associated with continuous changes in sonic characteristics.

Real-time interaction with virtual musical instruments provides another application domain. Kritsis et al. used LSTM-based models together with a Leap Motion sensor to recognize hand gestures for interaction with three-dimensional virtual musical instruments (Kritsis et al. 2018). This example demonstrates the importance of temporal modelling for gesture recognition, since meaningful interaction gestures unfold over time rather than appearing as isolated frames. It also illustrates a broader distinction between gesture recognition and embodied musical performance analysis: in interactive systems, the objective may be to recognize an intentional control gesture, whereas in performance analysis, the objective may be to understand how physical action contributes to the production and interpretation of musical sound.

Rhodes et al. compared machine learning approaches for classifying gestures within music and examined the use of different interfaces and computational methods for gesture-based musical interaction (Rhodes et al. 2020). Their work further demonstrates the diversity of gesture-analysis objectives, ranging from recognition and classification to real-time interaction. Such studies indicate that the term "musical gesture" encompasses both actions directly involved in sound production and movements used to control or communicate with computational musical systems.

More recent work has also extended gesture analysis toward video-based and multimodal performance analysis. Liwicki et al. investigated deep neural network approaches for analysing videos of music performances, demonstrating the potential of learned visual representations for extracting performance-related information from recorded musical activity (Simistira Liwicki et al. 2022). Although video-based approaches do not directly provide acoustic evidence of gesture, they are relevant to embodied musical analysis because they provide complementary information about posture, movement, and performer behaviour. Their integration with audio representations offers a pathway toward multimodal models in which physical action and acoustic outcome can be studied jointly.

Taken together, these case studies demonstrate that embodied musical gesture analysis encompasses several related but distinct analytical objectives. Audio-centered approaches attempt to infer gesture-related information from acoustic consequences, motion-centered approaches directly characterize physical action, and multimodal approaches seek to model relationships between movement and sound. In parallel, some studies focus on discrete gesture recognition, whereas others model continuous gesture–sound mappings or analyze expressive performance characteristics. This diversity highlights why no single representation or computational method is sufficient for embodied musical gesture analysis. It also motivates the need for evaluation strategies that distinguish recognition performance

from evidence that a learned representation captures transferable relationships between physical action and musical sound.

5.9. Representative Case Studies of Musical Gesture Analysis

The relationship between musical gesture and sound has been investigated through a range of complementary approaches, from indirect estimation of instrumental actions from acoustic characteristics to multimodal recognition using synchronized audio and motion recordings. These studies illustrate that musical gesture analysis is not a single computational task; rather, it encompasses gesture estimation, classification, sound–movement interaction, gesture–sound mapping, and real-time musical interaction. Table 3 summarizes representative studies according to the musical context, gesture being investigated, input representations, and analytical objective. The examples also illustrate an important distinction between approaches that infer gesture-related information from audio and approaches that directly measure physical movement. This distinction is central to embodied musical gesture analysis because acoustic information provides an indirect trace of physical action, whereas motion-based and multimodal recordings provide more direct measurements of the underlying bodily processes.

Several patterns emerge from these case studies. First, sound-producing gestures such as bowing, articulation, and instrumental interaction tend to exhibit stronger and more consistent acoustic correlates than expressive or communicative movements. Second, temporal representations are repeatedly important because gesture is inherently dynamic and its relationship with sound unfolds over time. Third, multimodal studies demonstrate that combining audio with motion can substantially improve gesture recognition, but such results should not be interpreted as evidence that audio alone fully represents the underlying physical action. Finally, the progression from handcrafted descriptors to machine-learned representations reflects a shift from explicitly interpretable acoustic correlates toward distributed representations that may capture more complex gesture–sound relationships. These observations motivate the subsequent discussion of audio descriptors and deep learning approaches, while also highlighting the need to evaluate whether learned representations capture transferable information about embodied performance rather than dataset-specific acoustic correlations.

6. Evaluation Methodologies for Embodied Audio Representations

Determining whether audio representations encode embodied information requires evaluation methods that test whether gesture-related and performance variables can be inferred from learned embeddings. Common approaches include probing representations to predict motion trajectories, expressive performance parameters, performer characteristics, or gesture categories (Dalmazzo and Ramirez 2019, 2021; Yu et al. 2022). However, successful prediction alone does not necessarily demonstrate embodiment, since models may exploit acoustic correlations rather than meaningful relationships between sound, movement, and physical action.

A central challenge is therefore distinguishing acoustic association from embodied representation. Tasks such as predicting instrument identity, performer style, or musical genre may reveal information correlated with performance behavior but do not directly assess whether representations encode gesture-related structure. More direct evaluation strategies include audio-to-motion prediction, cross-modal retrieval between audio and movement embeddings, gesture-conditioned generation, and analysis of shared latent spaces between acoustic and movement representations. Audio-to-motion prediction studies demonstrate that temporal relationships between musical sound and human movement can be computationally learned, suggesting that acoustic representations contain information related

to body dynamics (Shlizerman et al. 2017). Similarly, gesture-oriented representations based on body and finger keypoints provide a framework for evaluating whether learned representations correspond to structured performer movements rather than only acoustic categories (Gan et al. 2020).

A critical distinction must be made between gesture identification and embodied representation. Successful gesture classification demonstrates that a model can discriminate acoustic patterns associated with different actions, but it does not necessarily indicate that the representation captures the physical processes underlying sound production. Therefore, gesture encoding should be understood as the presence of structured action-related information within audio representations, rather than evidence of causal understanding or embodied cognition.

Reported performance metrics across gesture-recognition studies should be interpreted with caution. Prediction accuracy, retrieval performance, and reconstruction quality cannot be directly compared across datasets because they reflect differences in instruments, gesture definitions, annotation strategies, recording conditions, performer populations, and levels of acoustic–gesture coupling. Instruments with strong mechanical relationships between physical action and acoustic output may allow more reliable gesture inference than instruments where these relationships are indirect. Consequently, current results should be interpreted primarily as evidence of task-specific information encoded within representations rather than universal measures of embodied representation quality.

Another important challenge concerns generalization across performers, instruments, and musical contexts. Many existing evaluations are conducted within restricted domains, where training and testing conditions share similar acoustic and performance characteristics. Such settings may overestimate the extent to which learned representations capture general principles of musical action rather than dataset-specific acoustic patterns. Future benchmarks should therefore examine transfer across instruments, performance techniques, performers, and cultural contexts. This issue is particularly important because pretrained audio representations may inherit biases from their training data, resulting in embeddings that reflect dataset-specific correlations rather than transferable musical concepts (Wang et al. 2023).

Research in embodied music cognition provides complementary perspectives by emphasizing relationships between auditory perception, bodily experience, and interaction. Roddy and Furlong (2013) discuss embodied schemata as a framework for understanding how listeners interpret sound through bodily experience, while Deweppe and Leman (2011) highlight the importance of meaningful relationships between bodily actions and sonic outcomes in interactive musical systems. Similarly, Antle et al. (2009a, 2009b) demonstrate that embodied mappings influence perceived naturalness, usability, and interpretability in interactive audio environments. These perspectives suggest that computational evaluation should consider not only predictive performance but also whether learned relationships correspond to meaningful human interpretations of musical action.

For machine-learned audio representations, evaluation should consider multiple levels of embodiment. Sound-producing gestures, such as articulation, bowing, or striking, are expected to leave stronger acoustic traces than expressive or communicative movements. Therefore, benchmarks should evaluate not only gesture categories but also continuous movement variables, expressive parameters, performer-specific characteristics, and collective coordination patterns. Cross-modal representation-learning approaches based on shared embedding spaces and contrastive objectives provide additional opportunities for assessing whether acoustic and movement modalities preserve common semantic structures (Manco et al. 2022). Furthermore, studies examining gesture–melody correspondence

demonstrate that relationships between musical structure and movement can be computationally analyzed beyond simple classification tasks (Nadkarni et al. 2023).

Robust evaluation protocols should additionally account for confounding factors, including performer identity, instrument type, recording conditions, and musical context. Computational measures such as prediction accuracy, retrieval performance, and representation similarity should be complemented by human-centered evaluation of perceptual relevance, interpretability, and interaction quality. These combined approaches are necessary to determine whether learned audio embeddings capture meaningful embodied structures rather than merely reflecting statistical regularities in acoustic data.

Ultimately, the central question is not only whether machine-learned audio representations contain gesture-related information, but what forms of embodied information they preserve and under which conditions this information remains accessible. Current evidence suggests that learned representations encode partial and task-dependent traces of musical action, particularly for gestures strongly coupled to sound production. However, determining whether these traces correspond to meaningful embodied representations requires evaluation frameworks that move beyond accuracy-based benchmarks toward cross-modal, interpretable, and perceptually grounded analyses.

7. Key Contributions, Research Challenges, Limitations, and Future Research Directions

The review presented in this paper contributes to the emerging study of embodied musical gesture by examining whether and to what extent machine-learned audio representations preserve information related to performer action. Rather than treating gesture recognition accuracy as direct evidence of embodied representation, the review establishes a distinction between acoustic correlates of physical action and representations that meaningfully characterize embodied musical processes. This distinction provides a conceptual basis for evaluating audio representations according to the degree and type of embodiment that they preserve.

7.1. Key Contributions

The first contribution of this review is the development of a conceptual framework that organizes embodied musical information according to its relationship with sound production. The framework distinguishes between sound-producing gestures, such as bowing, striking, plucking, and fingering, and more indirect expressive or communicative movements, such as posture, body sway, conducting, and interpersonal coordination. This distinction is important because the acoustic observability of these gesture categories differs substantially. Sound-producing gestures generally produce measurable acoustic consequences, whereas expressive and communicative gestures may only be weakly or indirectly reflected in the audio signal.

Second, the review connects conventional acoustic descriptors with contemporary learned audio representations. Spectral, temporal, energy, pitch, and rhythmic descriptors provide interpretable acoustic correlates of performance actions, while deep and self-supervised representations provide mechanisms for learning distributed and potentially more complex relationships within musical audio. The review therefore situates the transition from handcrafted features to learned representations within an embodied perspective rather than treating representation learning solely as a problem of improving classification performance.

Third, the review highlights the importance of multimodal and cross-modal learning for studying embodied information. Audio provides a scalable but indirect representation of performance, whereas motion capture, video, inertial sensors, and related modalities

provide more direct measurements of physical action. The combination of these modalities offers a methodological framework for determining which aspects of movement are reflected in sound and which aspects cannot be reliably recovered from audio alone.

Fourth, the review identifies generalization and confounding factors as central evaluation issues. High performance within a restricted dataset may reflect performer identity, instrument characteristics, recording conditions, or musical material rather than transferable information about physical action. Consequently, the review argues that embodied audio representations should be evaluated across performers, instruments, recording environments, and musical contexts whenever possible.

Finally, the review proposes that embodied information should be understood as a continuum rather than a binary property of an audio representation. Representations may preserve strong information about acoustically coupled motor actions while containing weaker or more ambiguous information about expressive, communicative, or collective bodily processes. This perspective provides a more precise basis for interpreting evidence from existing gesture-recognition and representation-learning studies.

7.2. Research Challenges

Several challenges currently limit progress in embodied audio representation learning. The first is the fundamental non-uniqueness of the gesture–sound relationship. A single acoustic outcome can result from multiple physical actions, while the same physical action can generate different acoustic outcomes depending on the instrument, performer, environment, and performance context. This makes the inverse inference of physical action from audio inherently ambiguous.

A second challenge is the limited availability of synchronized and diverse audio–gesture datasets. Existing multimodal datasets provide valuable measurements of physical performance but are generally much smaller than large-scale audio corpora. They also vary considerably in instrument coverage, gesture definitions, performer populations, motion-capture technologies, and annotation strategies. This limits the ability to train general-purpose embodied representations and makes direct comparison between existing studies difficult.

A third challenge concerns disentangling embodied information from acoustic confounds. Instrument identity, performer characteristics, recording conditions, room acoustics, microphone placement, musical style, and repertoire can all influence the acoustic signal. A learned representation may therefore contain information that is predictive of a gesture without capturing a generalizable relationship between physical action and sound. Developing evaluation protocols that explicitly control or remove these confounding factors remains an important challenge.

A fourth challenge concerns interpretability. Deep and self-supervised representations distribute information across multiple latent dimensions, making it difficult to determine which aspects of the representation correspond to articulation, movement, timing, dynamics, or other performance characteristics. Explainability methods can identify influential acoustic regions or latent components, but these analyses do not by themselves establish that a model has learned the underlying physical mechanism.

Finally, there is a methodological challenge in defining what should constitute evidence of embodiment. Gesture classification, audio-to-motion prediction, cross-modal retrieval, reconstruction, and representation similarity each measure different properties. No single metric can therefore establish that a representation captures embodied musical information. More explicit evaluation criteria are needed to distinguish acoustic prediction from meaningful representation of physical action.

7.3. Limitations of the Current Literature and This Review

The existing literature has several limitations that should be considered when interpreting the conclusions of this review. First, much of the available evidence concerns sound-producing gestures and instrument-specific tasks, particularly violin bowing and other actions with relatively strong acoustic consequences. Evidence concerning expressive body movement, communicative gestures, interpersonal coordination, and collective embodiment remains comparatively limited. The conclusions of the review should therefore not be generalized equally across all forms of musical embodiment.

Second, existing studies use heterogeneous datasets and evaluation protocols. Differences in gesture definitions, recording conditions, participant populations, modalities, feature representations, model architectures, and train–test splits make quantitative comparison between reported results difficult. In particular, high recognition accuracies reported in different studies should not be interpreted as directly comparable measures of embodied representation quality.

Third, many studies rely on relatively restricted experimental settings. Within-performer or within-instrument evaluation may produce substantially easier prediction problems than cross-performer or cross-instrument evaluation. Consequently, some reported results may overestimate the extent to which learned representations capture general properties of musical action.

The review itself also has limitations. The literature spans several overlapping fields, including music cognition, music information retrieval, human–computer interaction, gesture recognition, multimodal learning, and representation learning. As a result, terminology and methodological assumptions vary considerably across studies. The structured narrative approach adopted here provides a framework for integrating these perspectives, but it does not constitute a formal quantitative meta-analysis. In addition, the rapid development of self-supervised and multimodal learning means that the technical landscape is changing quickly, and conclusions concerning particular architectures may evolve as larger datasets and newer representation-learning methods become available.

7.4. Future Research Directions

Future research should first prioritize the development of larger and more diverse synchronized audio–gesture datasets. Such datasets should include multiple performers, instruments, performance techniques, recording environments, and musical contexts, together with accurately aligned audio and movement measurements. Standardized metadata and annotation protocols would further facilitate comparison between studies and support reproducible benchmarking.

Second, future benchmarks should place greater emphasis on cross-domain generalization. Models should be evaluated using held-out performers, instruments, recording environments, and musical materials rather than relying primarily on random train–test splits. Such evaluation would help determine whether learned representations capture general relationships between physical action and sound or merely memorize dataset-specific acoustic characteristics.

Third, future work should investigate continuous movement and performance variables in addition to discrete gesture categories. Predicting trajectories of hand position, bow velocity, articulation, timing, dynamics, or other continuous variables can provide a more detailed assessment of whether audio representations preserve the temporal structure of physical performance. This would also better reflect the continuous nature of embodied musical action.

Fourth, cross-modal representation learning represents a promising direction for addressing the scale–fidelity trade-off between large audio datasets and smaller motion datasets. Audio and motion encoders could be trained jointly or aligned through contrastive, distillation, or shared embedding objectives. A particularly important research question is whether an audio encoder pretrained with multimodal supervision can retain gesture-related information when motion input is removed at inference time.

Fifth, future studies should investigate the internal structure of learned audio representations. Layer-wise probing, representation similarity analysis, controlled perturbations, and explainability methods could be combined to determine which levels of a representation retain information about timing, articulation, timbre, and performer action. Such analyses may clarify whether gesture-related information is concentrated in particular layers or distributed throughout the representation.

Sixth, research should move toward causal and intervention-based evaluation where feasible. Instead of asking only whether an embedding predicts a gesture, experiments could examine whether controlled changes in physical action produce corresponding and interpretable changes in the learned representation or its acoustic input. This would provide stronger evidence for meaningful gesture–sound relationships than observational prediction alone.

Finally, future research should incorporate perceptual and human-centered evaluation. Embodied musical meaning is not solely a property of measurable movement or acoustic features; it also involves how performers and listeners perceive relationships between action, sound, and expressive intention. Combining computational evaluation with perceptual experiments could therefore establish whether learned representations correspond to musically meaningful aspects of embodiment rather than only statistically predictable acoustic patterns.

Overall, the next stage of embodied audio representation research should move from demonstrating that gesture-related variables are predictable toward determining why they are predictable, which aspects of physical action are preserved, and whether those relationships transfer across performers, instruments, and contexts. Such a shift would help distinguish representations that merely exploit acoustic correlations from representations that provide a more faithful computational account of the relationship between human action and musical sound.

8. Conclusions

This review examined whether and to what extent machine-learned audio representations encode embodied gestural information underlying musical performance. By integrating perspectives from embodied music cognition, computational music analysis, and multimodal representation learning, this review focuses on the relationship between embodied musical processes and learned audio representations, extending beyond previous work that primarily addressed embodied perception, expressive performance modelling, or general music machine learning.

Musical embodiment was considered as a multi-level phenomenon involving sound-producing gestures, expressive movements, individual performance characteristics, and collective coordination. These forms of embodied information differ in their acoustic observability, abstraction level, and relationship to sound production. The framework presented in this review organizes embodied musical information according to these dimensions, highlighting that some aspects of performance are strongly reflected in acoustic signals, whereas others require multimodal evidence or remain difficult to infer from audio alone. Current evidence suggests that learned audio representations preserve partial and task-dependent traces of embodied musical information, particularly for gestures with strong

acoustic consequences, such as articulation, timing variation, and instrument-specific performance actions. However, higher-level expressive and communicative movements remain substantially underexplored because their relationships with sound are weaker, indirect, or context-dependent. Therefore, the presence of gesture-related information within an audio representation should not be interpreted as evidence that a model captures the physical mechanisms underlying musical action.

The analysis of existing datasets revealed a fundamental limitation in current research. Resources containing explicit measurements of performer movement provide strong evidence for embodiment but remain restricted in scale, duration, performer diversity, and instrument coverage. Conversely, large-scale audio datasets enable powerful representation learning but provide only indirect access to the physical actions responsible for sound production. This imbalance creates a fundamental trade-off between embodiment fidelity and scalability, limiting systematic evaluation of whether learned representations generalize across instruments, performers, and performance contexts. The transition from handcrafted acoustic descriptors to deep neural and self-supervised representations has enabled models to capture increasingly complex temporal, spectral, and contextual properties of musical sound. Nevertheless, learned models may exploit correlations associated with performer identity, instrument characteristics, recording conditions, or musical style rather than encoding general relationships between physical action and sound production. Consequently, high recognition accuracy should be interpreted as evidence of task-specific information encoded within representations rather than as direct evidence of embodied representation.

Evaluation of embodied audio representations therefore requires approaches that extend beyond prediction accuracy. Cross-modal learning, representation analysis, explainability methods, perturbation-based testing, and perceptual evaluation are necessary to determine whether models rely on gesture-related acoustic consequences or dataset-specific correlations. Future benchmarks should evaluate multiple levels of embodiment, including continuous movement variables, expressive parameters, performer individuality, and collective coordination, while testing generalization across instruments, performers, and musical contexts.

An important conclusion from this review is that embodiment should not be considered a binary property of audio representations. Instead, embodied information exists along a continuum determined by the strength of the gesture–sound relationship, ranging from strongly acoustically coupled motor actions to more abstract expressive and perceptual processes. Different evaluation strategies are therefore required for different levels of embodiment, and no single recognition benchmark can fully characterize the extent to which a model represents embodied musical information. Overall, embodied audio representation learning remains an emerging research direction at the intersection of music cognition, computational performance analysis, and artificial intelligence. A central open challenge is determining whether future systems can move beyond detecting acoustic correlates of human action toward computational representations that more faithfully characterize the relationships between physical performance processes and musical sound.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3D	Three-Dimensional
AI	Artificial Intelligence
AIST++	AI-Supported Interactive System for Dance
CNN	Convolutional Neural Network
CNN-LSTM	Convolutional Neural Network–Long Short-Term Memory
COLA	Contrastive Language–Audio Learning
DEAP	Database for Emotion Analysis using Physiological Signals
EEG	Electroencephalography
EMG	Electromyography
FMA	Free Music Archive
GAN	Generative Adversarial Network
HMM	Hidden Markov Model
IMU	Inertial Measurement Unit
k-NN	k-Nearest Neighbours
LSTM	Long Short-Term Memory
MAD-EEG	Music and Auditory EEG Dataset
MAESTRO	MIDI and Audio Edited for Synchronous Tracks and Organization
MFCCs	Mel-Frequency Cepstral Coefficients
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
MMIPs	Multimodal Instrument Performances
MoCap	Motion Capture
NMED	Naturalistic Music EEG Dataset
NMED-H	Naturalistic Music EEG Dataset–Headphones
NMED-T	Naturalistic Music EEG Dataset–Tones
OpenMIIR	Open-Source Music Information Retrieval EEG Dataset
PMEmo	Perceived Music Emotion Dataset
RNN	Recurrent Neural Network
SMPL	Skinned Multi-Person Linear Model
SSL	Self-Supervised Learning
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
Transformer	Transformer Neural Network
URMP	University of Rochester Music Performance dataset
VAE	Variational Autoencoder

References

Anders, Christopher J., Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. 2019. Finding and removing Clever Hans: Using explanation methods to debug and improve deep models. *arXiv* arXiv:1912.11425.

Antle, Alissa N., Greg Corness, and Milena Droumeva. 2009a. What the body knows: Exploring the benefits of embodied interaction in technology design. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. New York: ACM.

Antle, Alissa N., Milena Droumeva, and Greg Corness. 2009b. Playing with the sound maker: Exploring the role of embodied interaction in musical interfaces. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. New York: ACM.

Bacher, Reinhard. 2020. Improving gesture recognition with machine learning: A comparison of traditional machine learning and deep learning. In *Proceedings of the 17th International Conference on Accelerator and Large Experimental Physics Control Systems (ICALEPCS'19)*. Geneva: JACoW Publishing, pp. 214–8. [CrossRef]

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*. Red Hook: Curran Associates Inc., vol. 33, pp. 12449–60.

Best, Paul, Jean Bresson, and Diemo Schwarz. 2018. Musical gesture recognition using machine learning and audio descriptors. In *Proceedings of the 16th International Conference on Content-Based Multimedia Indexing (CBMI)*. New York: IEEE.

Bevilacqua, Frédéric, Norbert Schnell, Nicolas H. Rasamimanana, Bruno Zamborlin, and Fabrice Guédy. 2011. Online gesture analysis and control of audio processing. In *Musical Robots and Interactive Multimodal Systems*. Volume 74 of *Springer Tracts in Advanced Robotics*. Berlin/Heidelber: Springer, pp. 127–42.

Bolton, Jamie. 2004. *Gestural Extraction from Musical Audio Signals*. Ph.D. thesis, Queen Mary, University of London, London, UK.

Camurri, Antonio, and Gualtiero Volpe. 2011. Multimodal analysis of expressive gesture in music performance. In *Musical Robots and Interactive Multimodal Systems*. Edited by Jorge Solis and Kia Ng. Volume 74 of *Springer Tracts in Advanced Robotics*. Berlin/Heidelber: Springer, pp. 47–66. [CrossRef]

Camurri, Antonio, Barbara Mazzarino, Matteo Ricchetti, Renee Timmers, and Gualtiero Volpe. 2003. Multimodal analysis of expressive gesture in music and dance performances. In *Gesture Workshop*. Berlin/Heidelber: Springer.

Cancino-Chacón, Carlos E., Maarten Grachten, Werner Goebl, and Gerhard Widmer. 2018. Computational models of expressive music performance: A comprehensive and critical review. *Frontiers in Digital Humanities* 5: 25. [CrossRef]

Cantisani, Giorgia, Gabriel Trégoat, S. Essid, and Gaël Richard. 2019. Mad-eeg: An eeg dataset for decoding auditory attention to a target instrument in polyphonic music. In *SMM*. Dataset: MAD-EEG. Singapore: ISCA. [CrossRef]

Caramiaux, Baptiste, and Atau Tanaka. 2013. Machine learning of musical gestures. In *New Interfaces for Musical Expression*. New York: ACM.

Caramiaux, Baptiste, Jules Françoise, Norbert Schnell, and Frédéric Bevilacqua. 2014. Mapping through listening. *Computer Music Journal* 38: 34–48. [CrossRef]

Chen, Lei, Sylvie Gibet, and Camille Marteau. 2018. Conduct: An expressive conducting gesture dataset for sound control. In *International Conference on Language Resources and Evaluation*. Dataset: CONDUCT. Paris: ELRA. [CrossRef]

Clayton, Martin, Preeti Rao, Nithya Shikarpur, Sujoy Roychowdhury, and Jin Li. 2022. Raga classification from vocal performances using multimodal analysis. In *International Society for Music Information Retrieval Conference*. Bengaluru: ISMIR.

Cramer, Jesse, Justin Wu, and Daniel Ellis. 2021. Unsupervised learning for music audio representation. In *ISMIR*. Bengaluru: ISMIR.

Dalmazzo, Davide, and Rafael Ramirez. 2019. Bow gesture recognition in violin performance using machine learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Porto Alegre: UFRGS.

Dalmazzo, Davide, and Rafael Ramirez. 2021. Applying deep learning techniques to violin bowing gesture recognition. *Frontiers in Psychology* 12: 575971.

Deweppe, Atau, and Marc Leman. 2011. Methodological approaches to embodied musical interaction. In *Musical Gestures: Sound, Movement, and Meaning*. Oxfordshire: Routledge.

Dobrian, Christopher. 2012. A method for computer characterization of "gesture" in musical improvisation. In *Proceedings of the International Computer Music Conference (ICMC)*. Ann Arbor: Michigan Publishing Services.

Françoise, Jules, Norbert Schnell, and Frédéric Bevilacqua. 2013. Gesture-based control of physical modeling sound synthesis: A mapping-by-demonstration approach. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. New York: ACM. [CrossRef]

Gan, Chuang, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. 2020. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New York: IEEE, pp. 10478–87. [CrossRef]

Gemmeke, Jort F., Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New York: IEEE, pp. 776–80. [CrossRef]

Godøy, Rolf Inge. 2003. Motor-mimetic music cognition. In *Leonardo*. Cambridge: MIT Press, vol. 36, pp. 317–19. [CrossRef]

Han, Bo, Yi Ren, Hao Peng, Teng Zhang, Zeyu Ling, Xiang Yin, and Feilin Han. 2023. Enchantdance: Unveiling the potential of music-driven dance movement. *arXiv* arXiv:2312.15946.

Haunschmid, Verena, Ethan Manilov, and Gerhard Widmer. 2020. audiolime: Listenable explanations using source separation. In *Proceedings of the 13th International Workshop on Machine Learning and Music*. New York: ACM.

Hawthorne, Curtis, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling factorized piano music modeling and generation with the maestro dataset. *arXiv* 1810.12247.

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Kwang-Huei Saladi, Anuroop Sriram, Jiatong Zhang, James Glass, Abdelrahman Mohamed, and Alexei Baevski. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 3451–60. [CrossRef]

Huang, Yu-Fen, Nikki Moran, Simon Coleman, Jonathan Kelly, Shun-Hwa Wei, Po-Yin Chen, Yuning Huang, Tsung-Ping Chen, Yu-Chia Kuo, Yu-Chi Wei, and et al. 2024. Mosa: Music motion with semantic annotation dataset for cross-modal music processing. In *IEEE/ACM Transactions on Audio Speech and Language Processing*. Dataset: MOSA. New York: IEEE. [CrossRef]

Humphrey, Eric J., Juan Bello, and Yann LeCun. 2012. Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. In *International Society for Music Information Retrieval Conference*. Bengaluru: ISMIR.

Jensenius, Alexander Refsum, Marcelo M. Wanderley, Rolf Inge Godøy, and Marc Leman. 2009. Musical gestures: Concepts and methods in research. In *Musical Gestures: Sound, Movement, and Meaning*. Edited by Rolf Inge Godøy and Marc Leman. Oxfordshire: Routledge, pp. 12–35. [CrossRef]

Kaneshiro, Blair, Duc T. Nguyen, Jacek Dmochowski, Anthony Norcia, and Jonathan Berger. 2019. Naturalistic music eeg dataset—Hindi (nmed-h) 2.0: New release and cross-dataset compatibility. In *ISMIR 2019*. Dataset: NMED-H 2.0. Bengaluru: ISMIR.

Kim, Jaehun, Julián Urbano, Cynthia C. S. Liem, and Alan Hanjalic. 2019. Are nearby neighbors relatives? testing deep music embeddings. *Frontiers in Applied Mathematics and Statistics* 5: 55. [CrossRef]

Kim, Yonghyun, Junhyung Park, Joonhyung Bae, Kirak Kim, Taegyun Kwon, Alexander Lerch, and Juhan Nam. 2025. PianoVAM: A multimodal piano performance dataset. *arXiv* arXiv:2509.08800.

Koelstra, Sander, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3: 18–31. [CrossRef]

Kozak, Mariusz, Kristian Nymoen, and Rolf Inge Godøy. 2011. Effects of spectral features of sound on gesture type and timing. In *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication (GW 2011)*. Volume 7206 of *Lecture Notes in Computer Science*. Berlin/Heidelber: Springer.

Kritsis, Kosmas, Aggelos Gkiokas, Maximos A. Kaliakatsos-Papakostas, Vassilis Katsouros, and Aggelos Pikrakis. 2018. Deployment of LSTMs for real-time hand gesture interaction of 3D virtual music instruments with a Leap Motion sensor. In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*. New York: IEEE.

Kyriakou, Theodoros, Andreas Aristidou, and Panayiotis Charalambous. 2025. Multi-modal instrument performances (mmip): A musical database. *Computer Graphics Forum (Print)* 44: e70025. [CrossRef]

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* 10: 1096. [CrossRef] [PubMed]

Le, Nhat, Thang Pham, Tuong Khanh Long Do, Erman Tjiputra, Quang D. Tran, and A. Nguyen. 2023. Music-driven group choreography. In *Computer Vision and Pattern Recognition*. Dataset: AIOZ-GDANCE. New York: IEEE. [CrossRef]

Leman, Marc. 2007. *Embodied Music Cognition and Mediation Technology*. Cambridge: MIT Press.

Leman, Marc, and Antonio Camurri. 2006. Understanding musical expressiveness using interactive multimedia platforms. *Musicae Scientiae* 10: 209–33. [CrossRef]

Lembke, Sven-Amin. 2020. Sound-gesture identification in real-world sounds of varying timbral complexity. In *2nd International Conference on Timbre*. Thessaloniki: Aristotle University of Thessaloniki.

Lemouton, Serge, Riccardo Borghesi, Sampo Haapamäki, Frédéric Bevilacqua, and Emmanuel Fléty. 2019. Following orchestra conductors: The idea open movement dataset. In *MOCO*. Dataset: IDEA Open Movement. New York: ACM. [CrossRef]

Li, Bochen, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. 2016. *Creating a Musical Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications*. Dataset: URMP. Rochester: URMP.

Li, Ronghui, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, and Xiuhua Li. 2022. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *IEEE International Conference on Computer Vision*. Dataset: FineDance. New York: IEEE. [CrossRef]

Li, Ruilong, Sha Yang, David A. Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. New York: IEEE, pp. 13401–12.

Liang, Jing. 2023. Harmonizing minds and machines: Survey on transformative power of machine learning in music. *Frontiers in Neurorobotics* 17: 1267561. [CrossRef] [PubMed]

Losorelli, Steven, Duc T. Nguyen, J. Dmochowski, and Blair Kaneshiro. 2017. Nmed-t: A tempo-focused dataset of cortical and behavioral responses to naturalistic music. In *International Society for Music Information Retrieval Conference*. Dataset: NMED-T. Bengaluru: ISMIR.

MacRitchie, Jennifer, Manuel Varlet, and Peter E. Keller. 2017. Embodied expression through entrainment and co-representation in musical ensemble performance. In *The Routledge Companion to Embodied Music Interaction*. Oxfordshire: Routledge.

Maes, Pieter-Jan. 2016. Sensorimotor grounding of musical embodiment and the role of prediction: A review. *Frontiers in Psychology* 7: 308. [CrossRef] [PubMed]

Maes, Pieter-Jan, Marc Leman, Caroline Palmer, and Marcelo M. Wanderley. 2014. Action-based effects on music perception. *Frontiers in Psychology* 4: 1008. [CrossRef] [PubMed]

Maestre, Esteban, Panos Papiotis, Marco Marchini, Quim Llimona, Oscar Mayor, Alfonso Pérez, and M. Wanderley. n.d. Online Access and Visualization of Enriched Multimodal Representations of Music Performance Recordings: The Quartet Dataset and the Repovizz System. Dataset: Quartet Dataset (Repovizz). Available online: https://api.semanticscholar.org/CorpusID:41615074 (accessed on 16 September 2026).

Manco, Ilaria, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive audio-language learning for music. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*. Bengaluru: ISMIR.

Miliaresi, Ioanna. 2025. *Digital Audio Processing Methods for Voice Pathology Detection*. Doctoral thesis, University of Piraeus, Piraeus, Greece.

Morgado, Pedro, Juan Meseguer-Brocal, and Perfecto Herrera. 2021. Learning audio-visual correspondences for music. In *ICASSP*. New York: IEEE.

Moysis, Lazaros, Lazaros Iliadis, Sotirios Sotiroudis, Achilles Boursianis, Maria S. Papadopoulou, Konstantinos-Iraklis D. Kokkinidis, Christos Volos, Panos Sarigiannidis, Spiridon Nikolaidis, and Sotirios K. Goudos. 2023. Music deep learning: Deep learning methods for music signal processing—A review of the state-of-the-art. *IEEE Access* 11: 17031–52. [CrossRef]

Nadkarni, Shreyas Manish, Sujoy Roychowdhury, Preeti Rao, and Martin Clayton. 2023. Exploring the correspondence of melodic contour with gesture in raga alap singing. In *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*. Bengaluru: ISMIR, pp. 21–28.

Papiotis, Panos, Marco Marchini, Alfonso Perez-Carrillo, and Esteban Maestre. 2014. Measuring Ensemble Interdependence in a String Quartet Through Analysis of Multidimensional Performance Data. *Frontiers in Psychology* 5: 963. [CrossRef] [PubMed]

Peeters, Geoffroy, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. 2011. The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America* 130: 2902–16. [CrossRef] [PubMed]

Pons, Jordi, T. Lidy, and Xavier Serra. 2016. Experimenting with musically motivated convolutional neural networks. In *International Conference on Content-Based Multimedia Indexing*. New York: IEEE.

Reybrouck, Mark, and Andrea Schiavio. 2024. Music performance as knowledge acquisition: A review and preliminary conceptual framework. *Frontiers in Psychology* 15: 1331806. [CrossRef] [PubMed]

Rhodes, Chris, Richard Allmendinger, and Ricardo Climent. 2020. New interfaces and approaches to machine learning when classifying gestures within music. *Entropy* 22: 1384. [CrossRef] [PubMed]

Roddy, Stephen, and David Furlong. 2013. Embodied cognition and music: Theoretical and practical perspectives. In *Music, Sound and Movement*. Cambridge: The MIT Press.

Schiavio, Andrea, Damiano Menin, and Jakub Matyja. 2015. Music in the flesh: Embodied simulation in musical understanding. *Psychomusicology: Music, Mind, and Brain* 24: 340–43. [CrossRef]

Shlizerman, Eli, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. 2017. Audio to body dynamics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York: IEEE.

Simistira Liwicki, Foteini, Richa Upadhyay, Prakash Chandra Chhipa, Killian Murphy, Federico Visi, Stefan Östersjö, and Marcus Liwicki. 2022. Deep neural network approaches for analysing videos of music performances. *arXiv* arXiv:2205.11232. [CrossRef]

Stober, Sebastian, Avital Sternin, A. Owen, and Jessica A. Grahn. 2015. Towards music imagery information retrieval: Introducing the openmiir dataset of eeg recordings from music perception and imagination. In *International Society for Music Information Retrieval Conference*. Dataset: OpenMIIR. Bengaluru: ISMIR.

Thickstun, John, Zaid Harchaoui, and Sham Kakade. 2017. Musicnet: A curated dataset for music research. In *Proceedings of the ISMIR*. Bengaluru: ISMIR.

Tindale, Adam R., Ajay Kapur, and George Tzanetakis. 2011. Training surrogate sensors in musical gesture acquisition systems. *IEEE Transactions on Multimedia* 13: 50–59. [CrossRef]

Traube, Caroline, Philippe Depalle, and Marcelo Wanderley. 2003. Indirect acquisition of instrumental gesture based on signal, physical and perceptual information. In *New Interfaces for Musical Expression*. New York: ACM.

Tsuchida, Shuhei, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *International Society for Music Information Retrieval Conference*. Dataset: AIST Dance Video Database. Bengaluru: ISMIR.

Wang, Changhong, Gaël Richard, and Brian McFee. 2023. Transfer learning and bias correction with pre-trained audio embeddings. *arXiv* arXiv:2307.10834.

Wang, Changhong, Vincent Lostanlen, and Mathieu Lagrange. 2023. Explainable audio classification of playing techniques with layer-wise relevance propagation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New York: IEEE, pp. 1–5. [CrossRef]

Wang, Ruocheng, Pei Xu, Haochen Shi, Elizabeth Schumann, and C. Karen Liu. 2024. Fürelise: Capturing and physically synthesizing hand motion of piano performance. In *ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia*. Dataset: FürElise. New York: ACM. [CrossRef]

Yoo, Seong Jong, Snehesh Shrestha, Irina Muresanu, and Cornelia Fermüller. 2024. Viopose: Violin performance 4d pose estimation by hierarchical audiovisual inference. In *IEEE Workshop/Winter Conference on Applications of Computer Vision*. Dataset: VioPose. New York: IEEE. [CrossRef]

Yu, Jiashuo, Junfu Pu, Ying Cheng, Rui Feng, and Ying Shan. 2022. Self-Supervised Learning of Music-Dance Representation through Explicit-Implicit Rhythm Synchronization. *arXiv* arXiv:2207.03190. https://arxiv.org/abs/2207.03190. [CrossRef]

Zbyszyński, Michael, Balandino Di Donato, and Atau Tanaka. 2019. Gesture-timbre space: Multidimensional feature mapping using machine learning and concatenative synthesis. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*. Berlin/Heidelber: Springer.

Zhang, Kejun, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. 2018. The pmemo dataset for music emotion recognition. In *International Conference on Multimedia Retrieval*. Dataset: PMEmo. New York: ACM. [CrossRef]

Zhu, Lingting, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New York: IEEE, pp. 10544–53. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.