Sounds from Mismatch: Sensorimotor Prediction Error as Sonic Material in Augmented Reality

Domenico Stefani*
Alberto Boem*
Luca Turchet
dom@domenicostefani.com
alberto.boem@unitn.it
luca.turchet@unitn.it
DISI - Department of Information Engineering and Computer Science
Trento, Italy

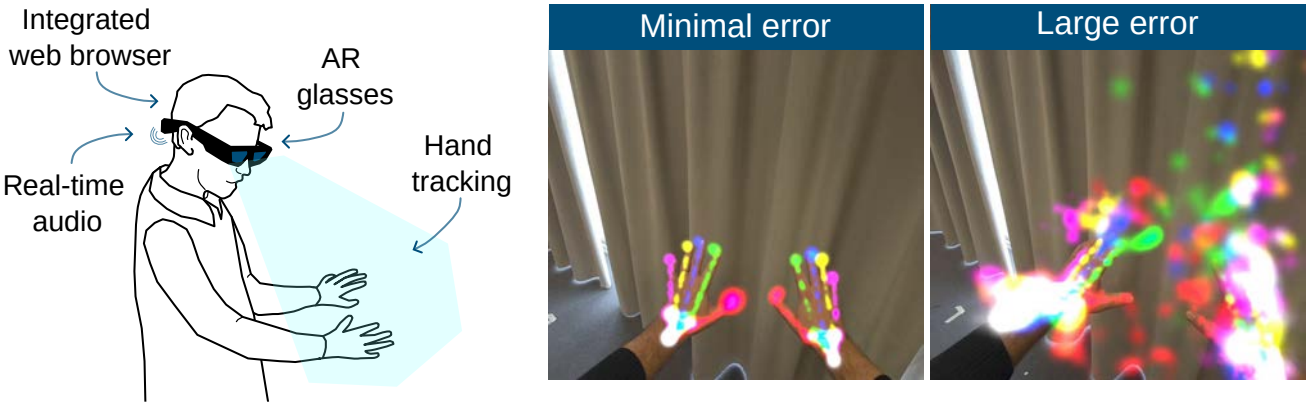

Figure 1: A visual summary of the components and the experience provided by Sounds from Mismatch. On the left, a user wearing a pair of AR glasses (it can also be a video passthrough head-mounted display) with hand-tracking, equipped with an integrated web browser that supports WebXR and Web Audio. On the right, two captures of the egocentric view of the user: one shows the visualization of a small prediction error, the other shows a visualization of a large prediction error.

Abstract

Sounds from Mismatch (SfM) is an augmented reality (AR) sound experience in which sound emerges from the mismatch between a user's bodily actions and the system's expectations of those actions. SfM builds an internal model of user interactions by maintaining individual predictors for future positions of multiple tracked hand joints. Rather than treating prediction error solely as a quantity to be minimized, SfM frames sensorimotor and expectation mismatch as an expressive material for interaction. Using hand tracking in AR, the system generates sound from the dynamic discrepancy between the user's movements and a continuously updated "ghost" representation of anticipated gestures. This mismatch is mapped to granular sound textures, producing an embodied auditory experience that unfolds through exploration rather than control. SfM invites users to attend to the shifting relationship between visual perception, proprioception, and action, foregrounding anticipation and deviation as core elements of musical interaction. This paper describes the conceptual framing, interaction design, and technical implementation of SfM, and discusses its implications for designing embodied AR sound experiences based on expectation rather than performative accuracy.

Keywords

Augmented Reality, Hand Tracking, Gesture-Sound Mapping, Predictive Models, Embodied Interaction

1 Introduction

In this paper, we propose Sounds from Mismatch (SfM): an augmented reality (AR) musical environment for see-through AR glasses and passthrough head mounted displays (HMDs). SfM approaches digital musical instruments not as a surface to be controlled, but as a predictive relation to be explored.

In conventional instrument design, the hand acts upon an external interface, and sound emerges from the success of that action, i.e., from the alignment between intention, gesture, and sonic outcome. In this context, mastery and virtuosity [32] involve reducing the gap between what the performer intends and what the instrument produces. Our work attempts to invert this logic: rather than rewarding accurate control, it generates sound from the degree of unpredictability of user actions that leads to failure of prediction. This is the discrepancy between what the user does and what the system anticipated they would do. The "instrument," if it can be called that, becomes not a physical object but a relational gap, the space between action and expectation. This shifts the site of musical interaction from the hand-based or hand-instrument boundary to what we might call the *hand-self*

*Both authors contributed equally to this research.

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

boundary: a pre-reflective, predictive sense of one's own body that normally operates invisibly beneath conscious awareness. Drawing on predictive processing accounts of cognition [42], we treat sensorimotor prediction error as a material to be sonified rather than reduced.

In SfM, the user encounters "ghost" hands: a visual representation of where the system predicts their hands should have moved based on recent history. When the actual hand diverges from this prediction, sound emerges. When they coincide, silence. The ghost is not a mirror of what the user has done, nor a representation of the present, but a projection of an anticipated future that did not occur, a strange temporal object that makes audible the normally transparent machinery of bodily self-modeling.

SfM goes back to the concept of "responsive environments" as proposed by Myron Krueger: "*Over a period of time the computer's displays establish a context within which the interaction occurs. It is within this context that the participant chooses his next action and anticipates the environment's response. If the response is unexpected, the environment has changed the context and the participant must reexamine his expectations. The experience is controlled by a composition which anticipates the participant's actions and flirts with his expectations*" [25]. In SfM, the "computer" maintains its own set of expectations, and its reactions are not designed to comfort the user, but to exhibit its predictive behavior. Users may not immediately understand how the environment responds to them, but this ambiguity invites exploration.

Unlike the idea of "responsive environments" envisioned by Krueger as public spaces, we reframe this concept as an egocentric experience. The sound is meaningful only to the person wearing a display equipped with tracking and processing capabilities (e.g., a pair of smartglasses or a head-mounted display), as a means of attending to their own sensorimotor dynamics. Users are not expected to play an "instrument" (as a surface to excite or a series of parameters to control), but to play against a model of themselves, exploring the boundaries of their own predictability, discovering where their body can surprise a system that is trying to anticipate it. In this way, SfM proposes a different relationship between body, prediction, and sound: one in which expression emerges not from control, but from the productive friction of being imperfectly modeled.

SfM was developed as a WebXR[1] application compatible with both novel consumer see-through AR glasses (i.e., Snap Spectacles[2]) and HMDs with video passthrough and hand-tracking capabilities. The source code is available as open-source[3] and the WebXR application is publicly hosted online[4].

Here, we present the background and conceptual framing behind SfM, the technical implementation of the system, and discuss its implications for the design of embodied AR sound experiences based on sensorimotor perception and predictive mismatch.

2 Background

2.1 Predictive Processing and Bodily Self-perception

Our proposed approach draws on the accounts of cognition made by predictive processing, which frame the brain as continuously generating predictions and updating them based on error signals [42]. The free energy principle of Friston [15, 36] or the work by Clark on predictive minds [9, 10] propose a description of perception as active hypothesis-testing. Seth extended this to bodily self-perception, arguing that our sense of our own bodies is itself a prediction - a "controlled hallucination" maintained by minimizing sensory prediction error, particularly from interoceptive and proprioceptive signals [40]. Our work draws inspiration from these concepts as SfM acts as a sort of a system that has a "brain" that maintains a model of user interactions and sonifies the "surprise" [15] defined as the mismatch between its model and reality.

2.2 Prediction in Musical Interfaces

Recent work has explored prediction in digital musical instruments. Notably, Martin et al. introduced EMPI [29], a system that uses recurrent neural networks to predict gestures of users. The predicted output drives different kind of actuators that responds to the performer, framing the prediction as enabling continuation [35] or call-and-response. Our work differs fundamentally: rather than using prediction output, we treat the error between prediction and action as the primary sonic material. In general, we stray from the concept of tactically predictive systems [45] (e.g., the predictive live-coding environment by Diapoulis et al. [11]). In SfM, sound represents where prediction fails, not what the system thinks you will do. A similar concept was hinted at by Pelinski et al. [37], who suggested devising models that predict the future behavior of sensor/control signals to create instruments that sonify the prediction error.

With respect to the works above, we chose to adopt algorithmic predictors instead of deep learning models as we aimed at running many predictors (i.e., one for each 3D axis of crucial tracked hand joints[5], of each tracked hand, resulting in 66 active predictors) in real-time on resource-constrained devices. While recent work [37, 43] showed promising results for integrating deep inference into resource-constrained devices, we found some of our target devices (i.e., Snap Spectacles) to offer much less computation headroom than others due to their compact size. This, paired with the large number of predictors required and the simpler nature of single-axis time series prediction, led us to adopt a range of algorithmic predictors.

2.3 The Performative Relation Between the Hand and the Self

NIME research has extensively studied the hand-instrument relationship, such as how hand gesture maps to sound [21, 28, 41, 46, 50]. Several studies were devoted to physical interfaces designed to augment musicians' hands (e.g., Waisvisz's "The Hands" [46], Sonami's "Lady's Gloves" [41]) and how such devices can be incorporated into performers' body schema [1, 33].

Moreover, in the context of AR and AR, hands occupy an important locus since they are used as the principal means for directly interacting with 3D user interfaces and virtual objects: from Lanier's paradigmatic performance "The Sound of One Hand" [27], relevant not only for its approach but also for its title itself, which points to the peculiarities of sensing and representing a singular body in virtual space, to recent explorations of hand-based interactions in the context of HMD-based AR (e.g., [2, 17, 18, 39, 47]). While Wang et al. [48] showed that free-hand interactions in AR can be easily learned in a musical context,

[1] https://github.com/immersive-web/webxr
[2] 5th generation Spectacles by Snap, Inc. with Snap OS, released to developers in 2024. https://www.spectacles.com/
[3] https://github.com/domenicostefani/sounds-from-mismatch-xr
[4] http://domenicostefani.com/sounds-from-mismatch-xr/
[5] i.e., wrist, metacarpals, and fingertips

it was also shown that due to tracking inconsistencies, this can't be a reliable solution in terms of timing precision [4].

However, these works focused on hands as input systems. Less explored is the hand-self relationship: the proprioceptive, predictive sense of one's own body before any external interaction. Merleau-Ponty described the body schema as a pre-reflective field of motor possibilities [30]; Gallagher distinguished this operative schema from explicit body image [16]. Work on the rubber hand illusion shows that bodily self-perception is constructed and easily disrupted [5]. Our work engages with an existing boundary between NIME and perception studies directly, using a predicted "ghost" hand to make audible the normally transparent process of bodily self-modeling.

Our work also connects to a long history of interactive art that makes the body an object of its own perception. In the landmark work Videoplace [26], Myron Krueger allowed users to see and interact with their silhouettes, creating what he termed an "artificial reality" through real-time responsive visual feedback. Such systems produce a doubled self - an image that is you but not quite you, responding but with its own logic. The ghost hand in our system functions similarly, but with a predictive projection: it reflects not what you are doing but what the system anticipated you would do. This makes the double predictive rather than mimetic, and the gap between self and double becomes the site of sonic interaction.

2.4 Egocentric Experience and Computing

The work discussed here departs from the implicitly performative orientation of most NIME work, where instruments are designed for expression directed outward. However, a thread that includes work on biofeedback music systems [44] and somaesthetic design [23], emphasizes inward-directed experience: interaction as attending to one's own body rather than communicating to others. This also resonates with recent research on wearable computing and smart glasses that prompted the use of first-person video and audio as the primary frame for understanding action, intention, and environment. For this, world models built on egocentric data aim to capture how a human agent perceives and predicts its own unfolding interaction with the world (e.g., [19, 20]).

Our work applies an analogous logic to sound-making and self-perception: the system models users' movements from their own embodied perspective, and the resulting experience is meaningful only from that first-person viewpoint. Where egocentric computing asks "*how does the world look from here?*", we ask "*how does my body feel from here - and what happens when that feeling is disrupted?*".

3 Implementation

The system was implemented using the WebXR Device API[6]. WebXR is an emerging web standard that brings virtual reality (VR) and AR experiences directly to the browser, eliminating the need for native applications. The standard provides a unified API that allows web applications to present immersive content using WebGL, handling camera settings and device interactions such as controllers, head, and hand tracking. Rather than requiring users to download native applications, WebXR brings immersive experiences directly to the web, making them accessible to anyone with a compatible device and a modern browser. WebXR has been explored in several works on musical instruments [3, 6, 8, 12, 13, 34].

[6]https://immersiveweb.dev/

We used the WebXR API through Three.js[7], which provides built-in support for immersive experiences. Three.js is a popular JavaScript library for developing interactive 3D graphics in the browser. Together, WebXR and Three.js handle stereoscopic rendering, camera tracking, and controller input automatically.

3.1 Visual Implementation

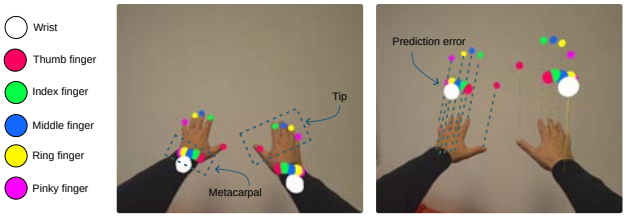

Figure 2: On the left, the visualization of the different joints used to debug the application. On the right, the prediction error of each joint is shown during a large movement of the hands.

Our application applies a particle visualization to the tracked joint of each hand[8]. We implemented a real-time hand visualization system using the WebXR hand-tracking functions[9] integrated with Three.js[10]. The system tracks 11 key skeletal joints per hand (wrist, metacarpals, and fingertips). For each, we compute prediction errors by comparing actual positions with forward predictions generated.

The visualization employs a particle-based rendering approach with three distinct particle systems: (1) zone particles that populate the volumetric space between actual and predicted joint positions, with particle density and opacity proportional to prediction error magnitude; (2) connection particles that form skeletal connections between anatomically-linked joints (palm structure, finger bones), creating a mesh-like hand representation; and (3) motion trails for legacy velocity visualization. All particles utilize custom GLSL shaders with additive blending for ethereal appearance and exponential fade-out. Joint-specific color coding (6-color palette) aids in visual differentiation of hand regions.

3.2 Audio Implementation

SfM sound implementation is based around two granular synthesizers, one for each hand. Granular synthesis was chosen as it grants a wide range of sonorities and quick prototyping by switching source material. SfM uses GrainPlayer from Tone.js[11]. GrainPlayer offers control over pitch, playback rate, grain size, overlap, and output volume. A mix of soundscapes and electronic sounds from Freesound[12] [14] was used as source material for the granular synthesizers.

Joint prediction errors were mapped to sound parameters by experimenting with a custom browser tool (Fig. 5, see repository[3]). The custom tool allows users to play back a recorded hand-tracking session and to see predictions and errors for all the different predictors and prediction horizons. This allowed

[7]https://threejs.org/
[8]https://developer.mozilla.org/en-US/docs/Web/API/XRHand
[9]WW3-WebXR Hand Input Module, https://www.w3.org/TR/webxr-hand-input-1/#feature-descriptor-hand-tracking
[10]Three.js WebXR Manager, https://threejs.org/docs/#XRManager
[11]https://tonejs.github.io/
[12]https://freesound.org

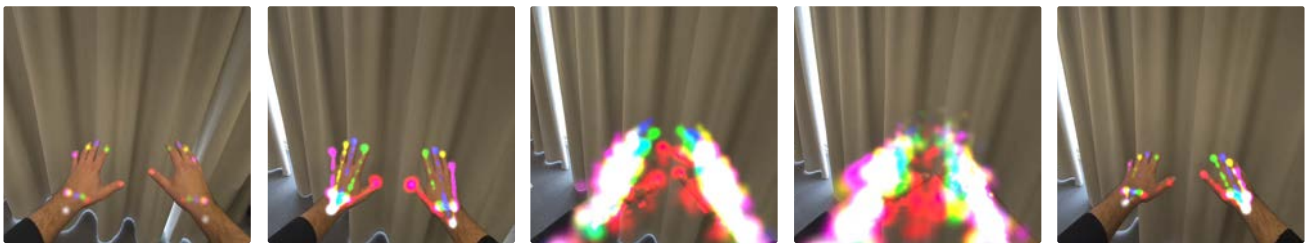

Figure 3: Ghost hand representation in the final version of SfM. From left to right, a sequence of actions showing the effects of the visualization of the prediction error.

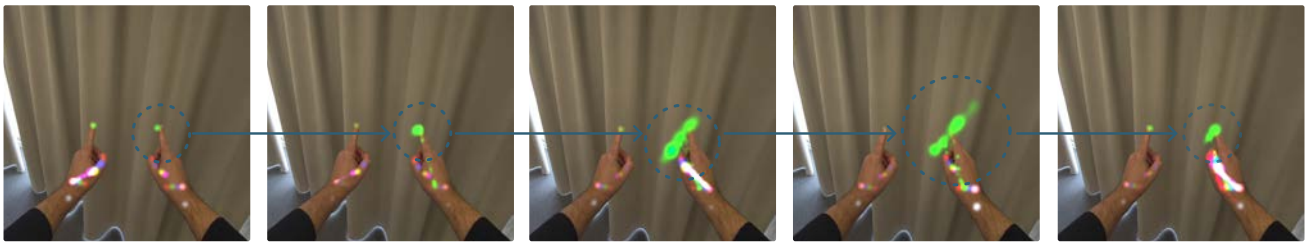

Figure 4: From left to right, a sequence of actions showing the effects of the visualization of the prediction error, specifically looking at the right-hand index finger.

tuning without repeatedly wearing a headset or AR glasses. Error-to-parameter mapping was made to gradually affect the sound timbre and volume with the increase of error. Mapping is discussed below.

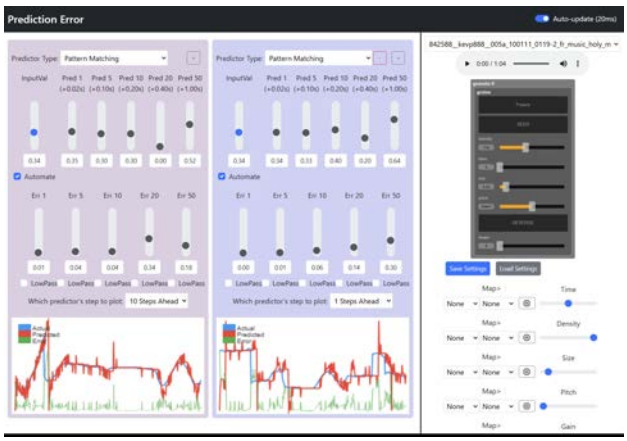

Figure 5: Custom error-sound mapping tool.

3.3 Predictors

For the aforementioned reasons concerning computational efficiency and time series prediction, we chose to adopt algorithmic predictors. The following predictors are employed:

- Linear Extrapolation (L) [22],
- Holt's Exponential Smoothing (H) [22],
- Kalman Filter (K) [24],
- Momentum (M) [31],
- Pattern Matching (P) [49],
- Sinusoidal Detection (S) [38].

For the sake of conciseness, formulations for each are not presented here. Implementations are found in the project repository[3]).

Predictors are attached to individual tracking signals (e.g., x-axis of the right-hand index finger tip) and executed at a 20ms refresh interval. Additionally, for each tracking signal, multiple predictors for different prediction horizons are run. At each refresh step, every predictor is fed the current sample from the relative tracking signal, and each is run for the number of steps of the relative prediction horizon (e.g., 1 step for 20ms, 50 steps for 1s). Each prediction result is appended to its relative queue buffer, which delays reading the prediction until its prediction horizon is reached in actual time. At the end, the prediction for the current actual time is extracted from every buffer and mapped to the visual *ghost* and the sound parameters. As a result, independently of the predictor/s chosen for the ghost-hand rendering, at any time, ghost-hand joints represent a past anticipation of the present, as it is a prediction for X refresh-intervals in the future, made X steps prior (X is the prediction horizon in refresh-intervals).

The performance of individual predictors was assessed on a 2-minute recording of hand-tracking data. The mean squared error (MSE) for each predictor and different prediction horizons is shown in Table 1 and Fig. 6. Despite the different performances, we chose to implement multiple predictors to have a wide palette of choices during mapping, which was informed by the quality of the shape of different errors in time (e.g., noise, repeating patterns, and slow adaptation in prediction output and error) and how these map to different parameters. This is why the aforementioned custom mapping tool was made to play actual handtracking signals and display plots of error trends. Pattern matching and sinusoidal detection were found to have low noise trends that could be mapped to sound parameters without additional filtering.

Furthermore, average error on the x and y axes of each thumb finger tip was mapped to the grain size, while average x,y error on index finger tips was mapped to pitch, and error on the x axis was mapped to volume. Error and parameter range mappings were tuned through trial and error.

Table 1: MSE of each predictor with different prediction horizons (Lowest per-horizon error in bold).

PHor*		+20ms	+100ms	+200ms	+400ms	+1s
Predictor**	L	3.9e-04	7.8e-04	1.7e-03	**4.6e-03**	2.7e-02
	H	2.5e-04	6.7e-04	1.9e-03	6.1e-03	3.0e-02
	K	6.6e-04	1.5e-03	2.9e-03	6.8e-03	3.0e-02
	M	3.4e-04	8.5e-04	1.9e-03	4.9e-03	2.7e-02
	P	**1.4e-04**	**5.5e-04**	**1.6e-03**	4.7e-03	**1.6e-02**
	S	1.9e-04	7.6e-04	2.0e-03	5.2e-03	**1.6e-02**

* *PHor: Prediction Horizon*

** *L: Linear Extrapolation, H: Holt's Exponential Smoothing, K: Kalman Filter, M: Momentum, P: Pattern Matching, S: Sinusoidal Detection*

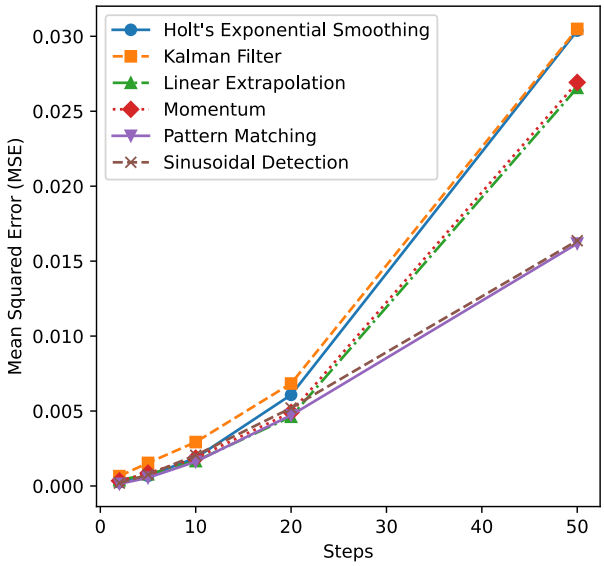

Figure 6: MSE of each predictor.

4 Design Considerations

Developing SfM surfaced several considerations on the design of both egocentric AR experiences and prediction-error-based sound mappings. We share reflections on visual and audio mappings, computational constraints, engagement modes, and their impact on egocentric AR-based musical systems.

4.1 Mapping

Visual Mapping. In our first experiences trying SfM before it had a visual rendering of the ghost hand, we found ourselves constantly going for a conventional control-based interaction, trying to position our hands in different ways and different parts of the space, often coming to a still state. Instead, a preliminary version, with a direct visualization of the prediction error (Fig. 2) elicited an opposite response: we found ourselves to enjoy the act of forcing prediction errors to increase through exaggerated movements of the hands and individual fingers. However, this limited the quality of the interaction to sudden movement bursts. When hands stayed still, predicted joint positions aligned with actual positions without visually matching the silence induced by the stillness (i.e., no prediction error occurred). This led us to devise a visual metaphor representing errors through blurred ghostly trails rather than direct error-to-visual mapping. When wearing the headset or glasses, the ghost hand creates a peculiar experience: you see where your hand was predicted to be a moment ago, creating a doubling that is neither synchronous nor simply delayed, but temporally twisted in a way that proves difficult to describe. Crucially, this representation gradually fades when approaching stillness, aligning with the principle that no prediction error equals no sound. However, this "blurred" rendering initially proves difficult to understand, and users benefit from an introduction to the system's concept and mechanics before use.

Sound Mapping. Differently, with respect to sound, mapping proved to be a more complex task. The first challenge arose from the limited independence of movements between joints in the same hand (i.e., small variations relative to the large x,y,z tracking ranges) that made differentiation difficult. However, granular synthesis provides a wide sound palette through just a few controls, like pitch and grain size. This constraint, combined with limited joint independence, led us to explore the use of joint prediction error along different axes.

To address varying error scales, we defined a dynamic and adaptive mapping range for the error values that updates in real-time based on encountered minimum and maximum values. This range shrinks over time to adapt to moments of exploration with different breadths of movement. The dynamic nature of the adaptive range mapping strategy means that decaying time is adjusted to leave time for wide ranges in between interactions, while preventing the system from going completely silent after a wide sweeping motion led to a large growth of the adaptive range. At the same time, it prevented the mapped value from saturating to a low fixed maximum, which otherwise hindered the dynamic range of sound parameters. Using a dynamic adaptive mapping came at the cost of having to adjust the decay time parameter finely, and experimenting with lowpassing error values for adaptive range calculation.

4.2 Technical Choices and Computational Constraints

In retrospect, technical constraints shaped the design of SfM at every stage. First, we chose algorithmic predictors over deep learning models based on initial assumptions about hardware limitations, particularly for AR glasses. Experimental testing then revealed further constraints that forced us to simplify our original design, especially regarding real-time sound generation and rendering. For instance, while our initial design employed a capable and complex granular processor[13] made in Faust and compiled as a Web audio module [7], we were forced to move to a lighter alternative in Tone.js to ensure smooth playback on Spectacles. Moreover, we had to reduce reverb processing as the combined computational cost of predictors, granular synthesis, and reverb exceeded the processing capacity of Spectacles. Similar issues were not found on the Meta Quest 3.

4.3 Engagement Modes

We had theorized that users would engage with SfM at a pre-reflective, sensorimotor level, but our own preliminary experiences during development suggest something more complex: moments of genuine proprioceptive confusion alternating with cognitive strategizing about how to "trick" the predictors.

[13] Granola Faust granular processor: https://github.com/jlp6k/faust-things/blob/main/Granola.dsp

Our first experiences with the system revealed moments where the ghost hand disrupted our perception of movement in disorienting yet enjoyable ways. This may be attributed to the behavior of the ghost hand, which was tightly coupled to the actual hand at times, and diverged at others, depending on the movements performed. For instance, we found that exploratory movements, rather than intentional gesture triggered these moments, making sound emerge from natural bodily dynamics. Importantly, we found that the experience felt more akin to discovering how one's own body behaves, or observing how a new coat moves semi-independently with the body, rather than playing an instrument.

However, this pre-reflective mode appeared difficult to sustain. Within minutes, interaction often seemed to shift toward more cognitive engagement: observing predictor behavior, identifying patterns in how the ghost responded to certain gestures, and deliberately exploiting these patterns. We found ourselves discovering that sudden direction changes confused predictors, or that repetitive movements would eventually be anticipated and produce silence. Whether this transition is inherent to the interaction paradigm or an artifact of our developer familiarity with the system remains an open question that structured user evaluation could address.

We also noticed that the prediction horizon appeared to affect this dynamic significantly. Shorter horizons (20-100ms) produced ghosts that stayed close to the actual hand, creating subtle, flickering discrepancies that felt more like perceptual noise. Longer horizons (400ms-1s) created ghosts that diverged more dramatically, making the prediction-reality gap more apparent. There may be an optimal range where the ghost is distant enough to be perceptually distinct but close enough to remain proprioceptively unsettling, though this hypothesis requires empirical validation. Sound itself seemed to play a role in modulating attention. When focusing on listening rather than watching, the visual ghost receded from attention, and interaction felt more exploratory. This suggests the multi-modal nature of the experience creates competing attentional demands.

These preliminary observations raise questions about designing systems that disrupt bodily self-perception: does exposing normally transparent processes to consciousness inevitably shift them from pre-reflective to reflective? Can this disruption persist, or does the system become absorbed into learned interaction patterns? Structured evaluation with participants unfamiliar with the system will help to better determine whether SfM achieves its intended effects or functions primarily as a cognitive puzzle. These studies should combine qualitative methods (e.g., phenomenological interviews, think-aloud protocols) with quantitative measurements (e.g., movement analysis, learning curves) to capture both embodied and cognitive dimensions of engagement.

5 Conclusion

We presented SfM as a preliminary experiment in reframing prediction error as primary sonic material rather than a quantity to minimize. Using hand tracking in AR, SfM generates sound from the mismatch between user movements and predicted "ghost" gestures, inverting conventional control paradigms that reward accuracy.

Developing SfM surfaced questions about prediction-based interaction that warrant further investigation: the tension between pre-reflective engagement and cognitive strategizing, the role of prediction horizon in phenomenological experience, and the interplay between visual and sonic attention in multimodal AR environments. Most critically, formal user evaluation is needed to determine whether the system achieves genuine body schema disruption or functions primarily as a cognitive puzzle.

Future work will combine phenomenological interviews with movement analysis to validate our preliminary observations. Furthermore, we aim to explore how such inherently egocentric experiences might be integrated into shared multiuser virtual environments. By presenting SfM at this preliminary stage, we hope to foster discussion about prediction, unpredictability as a sound parameter, and embodiment for this kind of egocentric AR experiences. We embrace NIME's welcoming of shorter-form contributions that can provoke similar kinds of novel conversations, even at early stages of development.

Ethical Standards

All exploratory testing conducted during the development of this system adhered to accepted principles of ethical research conduct. Participants engaged voluntarily, were informed about the nature of the system, and no personal data was collected or retained beyond the immediate testing context. The authors declare no conflicts of interest, financial or non-financial, related to the development or evaluation of SfM. The authors affirm that the research adheres to accepted principles of ethical and professional conduct within the NIME community, and no ethical issues were encountered in the conception, design, or reporting of this work.

Acknowledgments

We acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the European Innovation Council can be held responsible for them.

We thank Teresa Pelinski for the insightful chat on anomaly detection for gestural signals.

References

[1] Christopher Baber. 2003. *Cognition and tool use: Forms of engagement in human and animal use of tools.* CRC Press.

[2] Sam Bilbow. 2022. Evaluating polaris~ - An Audiovisual Augmented Reality Experience Built on Open-Source Hardware and Software. In *Proceedings of the International Conference on New Interfaces for Musical Expression.*

[3] Alberto Boem, Damian Dziwis, Matteo Tomasetti, Sascha Etezazi, and Luca Turchet. 2024. "It Takes Two" - Shared and Collaborative Virtual Musical Instruments in the Musical Metaverse. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2).* 1–10.

[4] Alberto Boem and Luca Turchet. 2024. Selection as Tapping: An evaluation of 3D input techniques for timing tasks in musical Virtual Reality. *International Journal of Human-Computer Studies* 185 (2024), 103231.

[5] Matthew Botvinick and Jonathan Cohen. 1998. Rubber hands 'feel' touch that eyes see. *Nature* 391, 6669 (1998), 756.

[6] Michel Buffa, Dorian Girard, and Ayoub Hofr. 2024. Using Web Audio Modules for Immersive Audio Collaboration in the Musical Metaverse. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2).* 1–10.

[7] Michel Buffa, Shihong Ren, Owen Campbell, Tom Burns, Steven Yi, Jari Kleimola, and Oliver Larkin. 2022. Web Audio Modules 2.0: An Open Web Audio Plugin Standard. In *Companion Proceedings of the Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22).* 364–369.

[8] Michel Buffa, Marco Winckler, and Adam Mir-Sadjadi. 2025. Embodied Virtual Instruments in Web-Based Multi-User VR: A Case Study with a 3D Drum Kit and Web Audio Modules. In *Proceedings of the 9th Web Audio Conference (WAC).* IRCAM/Mozilla.

[9] Andy Clark. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36, 3 (2013), 181–253.

[10] Andy Clark. 2015. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford University Press, New York.

[11] Georgios Diapoulis, Iannis Zannos, Kivanç Tatar, and Palle Dahlstedt. 2022. Bottom-up live coding: Analysis of continuous interactions towards predicting programming behaviours. In *Proceedings of the International Conference on New Interfaces for Musical Expression.* 15 pages.

[12] Damian Dziwis. 2023. Pdxr - The Evolution of Pure Data into the Metaverse. In *Proceedings of the International Computer Music Conference (ICMC).* 1–8.

[13] Damian Dziwis, Alberto Boem, Matthias Nowakowski, and Luca Turchet. 2025. Perspectives on Practical Implementations in the Web-based Musical Metaverse. In *2025 IEEE 6th International Symposium on the Internet of Sounds (IS2)*. 1–10.

[14] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*. 411–412.

[15] Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11, 2 (2010), 127–138.

[16] Shaun Gallagher. 2005. *How the Body Shapes the Mind.* Oxford University Press, Oxford.

[17] Max Graf and Mathieu Barthet. 2022. Mixed Reality Musical Interface: Exploring Ergonomics and Adaptive Hand Pose Recognition for Gestural Control. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Andrew McPherson and Emma Frid (Eds.). Article 44.

[18] Max Graf and Mathieu Barthet. 2023. Reducing Sensing Errors in a Mixed Reality Musical Instrument. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology* (Christchurch, New Zealand) *(VRST '23)*. Article 72, 2 pages.

[19] Kristen Grauman et al. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18995–19012.

[20] Kristen Grauman et al. 2024. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19383–19400.

[21] Yoonchang Han, Jinsoo Na, and Kyogu Lee. 2012. FutureGrab: A wearable subtractive synthesizer using hand gesture. In *Proceedings of the International Conference on New Interfaces for Musical Expression*.

[22] Charles C. Holt. 1957. *Forecasting Trends and Seasonal by Exponentially Weighted Averages.* Technical Report 52. Office of Naval Research. Reprinted in [need reprint details].

[23] Kristina Höök. 2018. *Designing with the Body: Somaesthetic Interaction Design.* The MIT Press, Cambridge, MA.

[24] R. E. Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering* 82, 1 (03 1960), 35–45.

[25] Myron W. Krueger. 1977. Responsive environments. In *Proceedings of the June 13-16, 1977, National Computer Conference* (Dallas, Texas) *(AFIPS '77)*. 423–433.

[26] Myron W. Krueger and Stephen Wilson. 1985. VIDEOPLACE: A Report from the ARTIFICIAL REALITY Laboratory. *Leonardo* 18, 3 (1985), 145–151.

[27] Jaron Lanier. [n. d.]. Virtual Reality and music. https://www.jaronlanier.com/vr.html

[28] James Leonard and Andrea Giomi. 2020. Towards an Interactive Model-Based Sonification of Hand Gesture for Dance Performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 369–374.

[29] Charles Patrick Martin, Kyrre Glette, Tønnes Frostad Nygaard, and Jim Torresen. 2020. Understanding Musical Predictions With an Embodied Interface for Musical Machine Learning. *Frontiers in Artificial Intelligence* Volume 3 - 2020 (2020), 14 pages.

[30] Maurice Merleau-Ponty. 1945. *Phenomenology of Perception.* Gallimard, Paris. Trans. Colin Smith, 1962. London: Routledge.

[31] Ali Mohammadi and Amin Kargarian. 2021. Momentum extrapolation prediction-based asynchronous distributed optimization for power systems. *Electric Power Systems Research* 196 (2021), 107193.

[32] Fabio Morreale, Andrew P. McPherson, and Marcelo Wanderley. 2018. NIME Identity from the Performer's Perspective. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 168–173.

[33] Luc Nijs, Micheline Lesaffre, and Marc Leman. 2009. The musical instrument as a natural extension of the musician. In *Proceedings of the 5th Conference of Interdisciplinary Musicology*. LAM-Institut jean Le Rond d'Alembert Paris, 132–133.

[34] Chaeryeong Oh, Dayoung Lee, and Alexandria Smith. 2024. Doongdoong.club: A Web-Based Metaverse Music Sequencer With Korean Onomatopoeic and Mimetic Words. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 622–626.

[35] François Pachet. 2003. The Continuator: Musical Interaction With Style. *Journal of New Music Research* 32, 3 (2003), 333–341.

[36] Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. 2022. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior.* MIT Press, Cambridge, MA.

[37] Teresa Pelinski, Rodrigo Diaz, Adan L. Benito Temprano, and Andrew McPherson. 2023. Pipeline for recording datasets and running neural networks on the Bela embedded hardware platform. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 160–166.

[38] Cameron N. Riviere, Robert S. Rader, and Nitish V. Thakor. 1998. Adaptive Cancelling of Physiological Tremor for Improved Precision in Microsurgery. *IEEE Transactions on Biomedical Engineering* 45, 7 (1998), 839–846.

[39] Giovanni Santini. 2020. Augmented Piano in Augmented Reality. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 411–415.

[40] Anil K. Seth. 2021. *Being You: A New Science of Consciousness.* Dutton.

[41] Laetitia Sonami. 2006. On my work. *Contemporary Music Review* 25, 5-6 (2006), 613–614.

[42] Mark Sprevak and Ryan Smith. 2023. An introduction to predictive processing models of perception and decision-making. *Topics in Cognitive Science* (2023).

[43] Domenico Stefani, Simone Peroni, and Luca Turchet. 2022. A Comparison of Deep Learning Inference Engines for Embedded Real-Time Audio Classification. In *Proceedings of the 25-th Int. Conf. on Digital Audio Effects (DAFx20in22)* (Vienna, Austria), Vol. 3. 256–263.

[44] Atau Tanaka and Marco Donnarumma. 2019. The Body as Musical Instrument. In *The Oxford Handbook of Music and the Body*. Oxford University Press, Oxford, Chapter 4.

[45] Steven L Tanimoto. 2013. A perspective on the evolution of live programming. In *2013 1st International Workshop on Live Programming (LIVE)*. IEEE, 31–34.

[46] Giuseppe Torre, Kristina Andersen, and Frank Baldé. 2016. The Hands: The Making of a Digital Musical Instrument. *Computer Music Journal* 40, 2 (2016), 22–34.

[47] Yichen Wang and Charles Martin. 2022. Cubing Sound: Designing a NIME for Head-mounted Augmented Reality. In *Proceedings of the International Conference on New Interfaces for Musical Expression*.

[48] Yichen Wang, Mingze Xi, Matt Adcock, and Charles Patrick Martin. 2023. Mobility, Space and Sound Activate Expressive Musical Experience in Augmented Reality. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 128–133.

[49] Andreas S Weigend. 2018. *Time series prediction: forecasting the future and understanding the past.* Routledge.

[50] Yue Yang, Zhaowen Wang, and Zijin Li. 2023. MuGeVI: A Multi-Functional Gesture-Controlled Virtual Instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Miguel Ortiz and Adnan Marquez-Borbon (Eds.). 536–541.