

# Vibro-Sense: Robust Vibration-based Impulse Response Localization and Trajectory Tracking for Robotic Hands

Wadhah Zai El Amri<sup>1\*</sup> and Nicolás Navarro-Guerrero<sup>1</sup>

<sup>1</sup>Leibniz Universität Hannover, L3S Research Center, Appelstraße 4, Hanover, 30167, Germany.

\*Corresponding author(s). E-mail(s): [wadhah.zai@l3s.de](mailto:wadhah.zai@l3s.de);  
Contributing authors: [nicolas.navarro.guerrero@gmail.com](mailto:nicolas.navarro.guerrero@gmail.com);

## Abstract

Rich contact perception is crucial for robotic manipulation, yet traditional tactile skins remain expensive and complex to integrate. This paper presents a scalable alternative: high-accuracy whole-body touch localization via vibro-acoustic sensing. By equipping a robotic hand with seven low-cost piezoelectric microphones and leveraging an Audio Spectrogram Transformer, we decode the vibrational signatures generated during physical interaction. Extensive evaluation across stationary and dynamic tasks reveals a localization error of under 5 mm in static conditions. Furthermore, our analysis highlights the distinct influence of material properties: stiff materials (e.g., metal) excel in impulse response localization due to sharp, high-bandwidth responses, whereas textured materials (e.g., wood) provide superior friction-based features for trajectory tracking. The system demonstrates robustness to the robot's own motion, maintaining effective tracking even during active operation. Our primary contribution is demonstrating that complex physical contact dynamics can be effectively decoded from simple vibrational signals, offering a viable pathway to widespread, affordable contact perception in robotics. To accelerate research, we provide our full datasets, models, and experimental setups as open-source resources.

**Keywords:** Tactile Sensing, Vibro-Acoustic Sensing, Contact Localization, Deep Learning

## 1 Introduction

Humans possess a comprehensive system for perceiving their physical surroundings that extends beyond exteroceptive senses like vision and hearing. The somatosensory system provides crucial information about direct contact and environmental forces Abraira and Ginty (2013); Navarro-Guerrero et al. (2023). When an object contacts with the body, the resulting mechanical vibrations propagate through the skin and tissues, enabling immediate localization and characterization of the interaction. This distributed vibro-acoustic sensing is fundamental not only for object manipulation

but for overall spatial awareness and safe interaction within a dynamic environment. The ability to detect and interpret these contact events is a key component of effective physical interaction Wang et al. (2021).

Drawing from this biological paradigm, the field of robot perception is increasingly exploring structure-borne sound analysis to supplement conventional vision and force modalities Bonner et al. (2021); Toprak et al. (2018). The integration of piezoelectric sensors, particularly contact microphones, presents a promising avenue for capturing the complex dynamics of robot-environment

interactions. By leveraging the unique properties of acoustic signals, researchers have made notable progress in developing vibro-acoustic sensing techniques for contact-rich manipulation Lu and Culbertson (2023); Mejia et al. (2024); Liu and Chen (2024); Wall and Brock (2022).

In this paper, we propose a cost-effective yet accurate method that enables robots to perceive physical contact in a more natural manner. We demonstrate our approach on two real-world tasks: impulse response localization and trajectory tracking. Our method utilizes a robotic hand equipped with seven contact microphones to capture vibrational signals, which are then processed using an Audio Spectrogram Transformer (AST) architecture to predict the positions of external touches on the hand. To train and evaluate this system, we collected extensive datasets consisting of over 65,000 unique samples for impulse response localization and over 240,000 interactions for trajectory tracking.

Our contribution is two-fold. Firstly, we present a robust method for vibro-acoustic sensing that can be applied to complex robotic geometries. We provide a detailed analysis of how material properties, specifically stiffness versus texture, distinctly influence localization accuracy across different interaction modes. Secondly, we demonstrate the effectiveness of our approach in dynamic scenarios, where the robot hand actively moves while interacting with its surroundings. Our results show that the system maintains effective localization accuracy even in the presence of significant motion and actuator noise. By leveraging vibro-acoustic sensing and deep learning, our work provides a scalable solution for whole-body contact perception. To facilitate further research, we open-source all our code, datasets, and experimental setups on our website, allowing researchers and practitioners to easily replicate and build upon our work: [wzaielamri.github.io/publication/vibrosense](https://wzaielamri.github.io/publication/vibrosense).

## 2 Related Works

For robots to operate safely and effectively in the physical world, they must be able to perceive and interpret contact within their environment. This sense of touch is crucial for enabling robots to securely manipulate objects and avoid unexpected or damaging collisions Hoffmann and Longo (2022). To address this, researchers have explored various



**Fig. 1** Setup of the impulse response localization task. The UR5e robotic arm is shown applying controlled pokes to the hand using a solenoid actuator with a metal indenter.

sensing modalities, each with distinct advantages and trade-offs. The dominant approaches include creating large-area robotic skin to directly mimic biological touch and developing high-resolution vision-based sensors for detailed, localized contact analysis Hardman et al. (2025). As a distinct alternative, vibro-acoustic sensing has emerged as a method for capturing the rich dynamics of interaction events, leveraging sound and vibration Lee et al. (2025). This section provides an overview of the state-of-the-art approaches, highlighting their respective capabilities and limitations.

### 2.1 Tactile Sensor Arrays and Robotic Skin

Tactile sensor arrays and robotic skin technologies enable robots and prosthetic devices to sense and interpret touch, mimicking how human skin detects pressure, texture, and temperature. These systems are built from grids of sensitive elements embedded in flexible materials, enabling accurate, responsive detection of physical contact and environmental interactions. An example is Touchlab, which develops ultra-thin electronic skin. Their sensors equip robots with real-time touch sensing, enabling remote control with haptic feedback for healthcare and hazardous environments. Hardman et al. Hardman et al. (2025) developed another advanced tactile skin solution that uses a soft hydrogel embedded with over 860,000 conductive pathways to detect pressure, temperature,

and damage, enabling highly sensitive, adaptive touch sensing for robots and prosthetics. However, these methods face significant drawbacks. They are often expensive to produce and implement, and their delicate construction makes them fragile and susceptible to damage from routine physical interaction. Furthermore, the sensing capability is localized, meaning it is confined only to the specific areas where the robotic skin is applied, limiting the robot’s overall environmental awareness Chen et al. (2023).

## 2.2 Vision-Based Tactile Sensing

Vision-based tactile sensing takes a different approach by inferring contact information from camera observations of a deformable material. This approach can reconstruct detailed 3D contact geometry and force distribution from images. Examples of such sensors include GelSight Yuan et al. (2017), which uses a camera to capture fine-grained texture and shapes of an illuminated elastomer. More compact designs such as DIGIT Lambeta et al. (2020) have helped standardize this approach and broaden its use in manipulation research. These sensors provide rich data well-suited for grasping and in-hand manipulation, but they remain inherently local. Due to their bulk and reliance on internal optics, they are inherently difficult to scale for large surface areas and are not well-suited for whole-body tactile sensing.

## 2.3 Vibro-Acoustic Sensing in Robotics

Distinct from surface-based electronic or visual methods, vibro-acoustic sensing exploits structure-borne sound to detect physical interactions. Contact microphones convert mechanical vibrations propagating through the robot body into electrical signals, providing a lightweight, inexpensive, and highly scalable sensing modality. It can provide a complementary source of information to vision and force sensors, especially in occluded settings Toprak et al. (2018). In the following, we review relevant microphone-based work on (i) impulse response localization, which estimates contact locations from transient vibration patterns and (ii) trajectory estimation during contact-rich tasks,

which tracks continuous motion from ongoing acoustic signals.

### 2.3.1 Impulse Response Localization

Instead of covering the robot with dense arrays, impulse localization estimates contact coordinates by decoding the structural vibrations captured by a sparse microphone array. For instance, SonicBoom Lee et al. (2025) equips a robot end-effector with a distributed array of piezoelectric contact microphones to estimate the 3D locations of contact events. Using a data-driven approach that leverages relative acoustic features across microphones, SonicBoom achieves centimeter-level localization accuracy across varying surfaces and contact types, demonstrating the potential of passive acoustic sensing for high-resolution spatial perception even in occluded environments.

While passive methods capture naturally occurring contact vibrations, active vibro-acoustic methods introduce controlled vibrations into the robot or object to probe contact. Lu and Culbertson Lu and Culbertson (2023) integrated piezoelectric actuators with microphones in robotic grippers, enabling closed-loop estimation of contact state and object properties from reflected acoustic signals. Similarly, Wall et al. Wall et al. (2023) embedded microphones into soft pneumatic actuators, enabling touch localization on deformable materials with complex geometries and in noisy environments. Such demonstrations highlight how acoustic sensing can augment contact perception without relying on dense tactile arrays or visual feedback.

Building on localization, the same vibration signals also reveal finer details, such as texture, slip, and material properties, making contact microphones highly complementary to other modalities. This richness enables biomimetic tactile designs, such as the fingerprint-inspired sensors from Quilachamín et al. Juiña Quilachamín and Navarro-Guerrero (2023) that enhance spatial resolution and material identification. These capabilities naturally extend to tracking continuous interactions.

### 2.3.2 Trajectory Tracking

Although locating discrete impacts is well-studied, extending vibro-acoustic sensing to track the continuous motion of sliding or drawing interactions

remains a significant challenge that has been addressed by fewer scientists. Liu and Chen Liu and Chen (2024) proposed SonicSense, employing in-hand acoustic vibration sensing combined with deep neural networks to reconstruct detailed object motion information during manipulation. By analyzing vibration-induced acoustic patterns generated by sliding or interacting objects on gel surfaces, their system estimates motion direction, speed, and reconstructs 3D shapes using sound as the primary sensing modality.

In a related direction, Lu et al. Lu et al. (2020) modeled and rendered tool-surface interaction sounds with wavelet-tree models segmented by contact velocity. This method, though primarily for perceptual rendering, demonstrates that velocity-dependent acoustic cues encode contact-motion information, underscoring the potential for trajectory inference from sound in manipulation tasks.

Extending this notion further, MilliSonic Wang and Gollakota (2019) achieves highly accurate acoustic motion tracking with sub-millimeter precision using airborne acoustic signals during free-space object motion. While this approach (i.e., airborne acoustic) lies outside the primary focus of this paper, it underscores the broader potential of acoustic modalities for precise trajectory estimation, even though it relies on external microphone arrays rather than contact-based vibration sensing.

These studies indicate that acoustic feedback contains rich spatiotemporal information to reconstruct continuous trajectories of hands or tools, opening new possibilities for audio-driven robotic control, high-resolution trajectory tracking, and dynamic, contact-rich manipulation.

The reviewed literature highlights that microphones can serve as lightweight and information-rich sensors for robotic interaction perception. However, the simultaneous application of this modality to both impulse response localization and continuous trajectory tracking, particularly while the robot itself is in motion, remains underexplored. Our work addresses this gap. In contrast to prior studies, we demonstrate that distributed contact microphones, coupled with a deep learning model, can robustly decode diverse contact behaviors. While our hardware setup shares the passive, multi-microphone design of *SonicBoom* Lee et al. (2025), we significantly extend the scope to include dynamic trajectory localization, tracking external

sliding contacts on the surface. Additionally, we provide a novel analysis of how material properties (e.g., stiffness vs. texture) distinctly influence the vibrational signatures used for localization. Furthermore, we validate our system’s robustness to the robot’s own actuator noise, a critical step toward enabling safe, whole-body contact perception in real-world deployment.

## 3 Method

We propose a cost-effective yet accurate method that enables robots to perceive physical contact in a more natural manner. We demonstrate our approach on two real-world tasks: impulse response localization and a more complex case, trajectory tracking.

The following section describes the hardware setup of the robotic hand, followed by a detailed explanation of the two tasks.

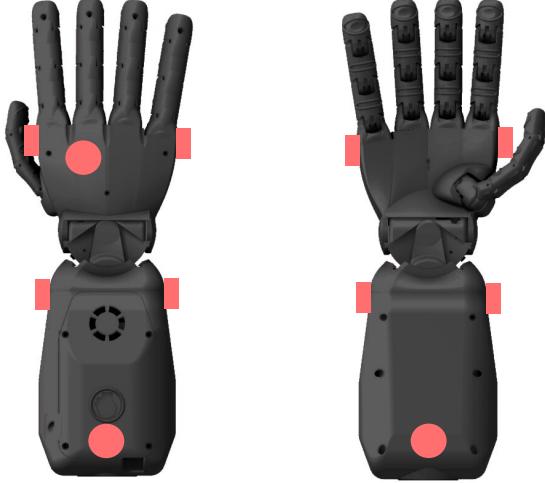
### 3.1 Hardware Setup

For all our experiments, we used the Seed Robotics’s RH8D hand Seed Robotics. This hand has 19 degrees of freedom and is actuated by 8 motors. We equip the hand with seven Harley Benton CM-1000 contact microphones positioned to capture tactile vibrations (Fig. 2). Custom mounts were designed to fix the sensors in position; corresponding CAD models are provided on the project website. These microphones are capturing a range of  $-500mV$  to  $+500mV$ .

### 3.2 Impulse Response Localization Task

#### 3.2.1 Task Overview

For the impulse response localization task, we employed a UR5e robotic arm Universal Robots equipped with a solenoid actuator, onto which we mounted four interchangeable cylindrical indenters made of distinct materials: soft plastic, hard plastic, wood, and metal, as shown in Figure 1. The UR5e autonomously moves the solenoid to randomized positions facing the ‘Back’, ‘Front’, ‘Right’, and ‘Left’ sides of the robotic hand. Once in position, the solenoid is activated to deliver controlled poking interactions to the surface of the hand. A demonstration video of this data collection process is available on our website.



**Fig. 2** Schematic diagram of the microphones' localization, represented by the red area. The microphones on the RH8D hand are mounted externally.

### 3.2.2 Dataset

The interactions generate mechanical vibrations and acoustic signals that propagate throughout the robotic hand's structure. Capturing these responses allows us to study the relationship between the resulting sensory feedback and the contact location. To build a robust dataset for training and evaluation, we collected approximately 65000 unique samples. Data were gathered from multiple sides and contact points around the robotic hand, both while the hand was idle and powered on and while it was off, to account for the background and internal noise introduced by the fan during normal operation.

Each interaction recording lasts 500 ms, triggered 200 ms prior to contact. We initially captured raw signals at 50 kHz to ensure full spectral coverage. To prepare this data for the neural network, we apply the following standardized preprocessing pipeline:

- 1. Downsampling:** The signals are downsampled to 20 kHz. As analyzed in Section 4.1, this sampling rate was selected because frequencies above this threshold were found to contain negligible information for localization.
- 2. Windowing:** We trim the recording to a focused 200 ms window (from 125 ms to 325 ms) to isolate the interaction event while discarding noise.

**3. Feature Extraction:** The processed time-series data is converted into time–frequency features using a Short-Time Fourier Transform (STFT). We use a window size ( $n_{fft}$ ) of 128, which was empirically determined to minimize localization error (see Section 4.1).

**4. Noise Removal:** We utilize the discarded initial 100 ms of the pre-trigger phase to estimate and subtract the steady-state background noise.

## 3.3 Trajectory Tracking Task

### 3.3.1 Task Overview

The trajectory tracking task is a more complex task, in which the UR5e robot arm draws different patterns and drawings on the surface of the forearm of the Seed Robotics's RH8D hand Seed Robotics using the different four indenter materials. To automate the data collection procedure, we program the UR5e arm to draw a subset of real drawings from the open-source *Quick Draw* Jongejan et al. (2016) dataset, which spans 345 categories. We use the simplified version of the drawings. To minimize the robot's travel time between strokes, we optimized the drawing sequence by formulating the problem as a Generalized Traveling Salesperson Problem (GTSP) Pop et al. (2024) and implemented this using the Google OR-Tools routing engine. This approach allows each stroke to be drawn either forwards or in reverse and adjust its drawing order.

### 3.3.2 Dataset

For this task, we collect two datasets with different levels of complexity. The first dataset consists of 40,000 strokes per indenter (160,000 in total), with the Seed robotic hand powered on, idling in a fixed position, and its fan noise present. The second dataset contains 20,000 interactions per indenter (80,000 in total), during which the Seed robotic hand moves into random positions. Each stroke interaction varies in duration between 1s and 10s. This setup mimics real-world conditions, albeit exaggerated here to test the approach's limits, in which the robot performs actions while still needing to correctly sense its surroundings. The second dataset is therefore more challenging, as it requires the network to distinguish between hand motion and external touches.

Preprocessing follows the steps described in Section 3.2.2. We downsample the signals to 20 kHz. We then segment the continuous signals into 200 ms chunks. For the target variable, we assign the average hand position ( $x, y, z$ ) recorded within each chunk. We then apply the same STFT parameters (128-window size) to generate consistent spectrogram features across both tasks. Finally, we remove the background noise using the first 100 ms of the recording.

### 3.4 Networks Architecture

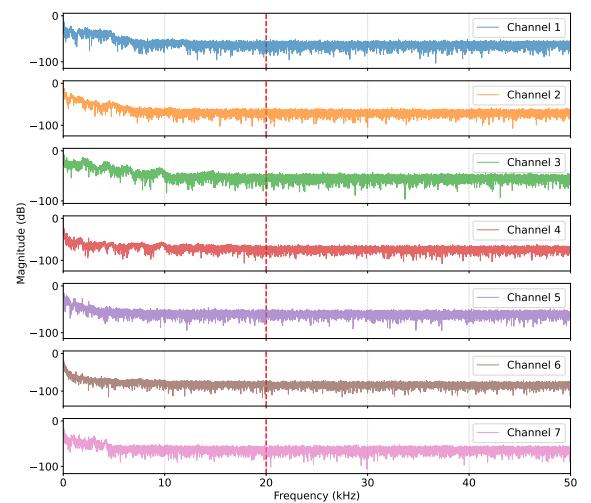
We employ the Audio Spectrogram Transformer (AST) Gong et al. (2021) architecture for these tasks due to its demonstrated effectiveness in learning from time–frequency representations of audio signals. AST leverages self-attention mechanisms to capture both local and global dependencies in spectrograms, making it particularly well-suited for tasks where temporal dynamics and subtle acoustic cues are critical. Its architecture allows for focusing on relevant patterns across the entire input, which is essential for accurately predicting target positions from audio interactions. Additionally, AST’s proven performance across a variety of audio classification and localization tasks ensures it generalizes well across our datasets, which vary in complexity.

While the standard AST architecture is designed for single-channel input, we adapt it for spatial perception by modifying the initial patch embedding layer to accept a 7-channel input tensor. We stack the synchronized spectrograms from the seven microphones along the channel dimension (creating a  $7 \times T \times F$  tensor, where  $T$  denotes the number of time frames and  $F$  the number of frequency bins), allowing the model to learn inter-channel spatial features, such as phase and amplitude differences, directly during tokenization. Our network configuration includes 12 transformer blocks with a kernel size of 16 and a stride of 10, and we use a batch size of 128. We optimize the model using mean squared error (MSE) loss, with the Adam optimizer and a learning rate of 0.0007. To improve stability, we apply a cosine learning rate schedule with a linear warmup phase (1% of the total training steps).

## 4 Experimental Results

### 4.1 Parameter Optimization & Spectral Analysis

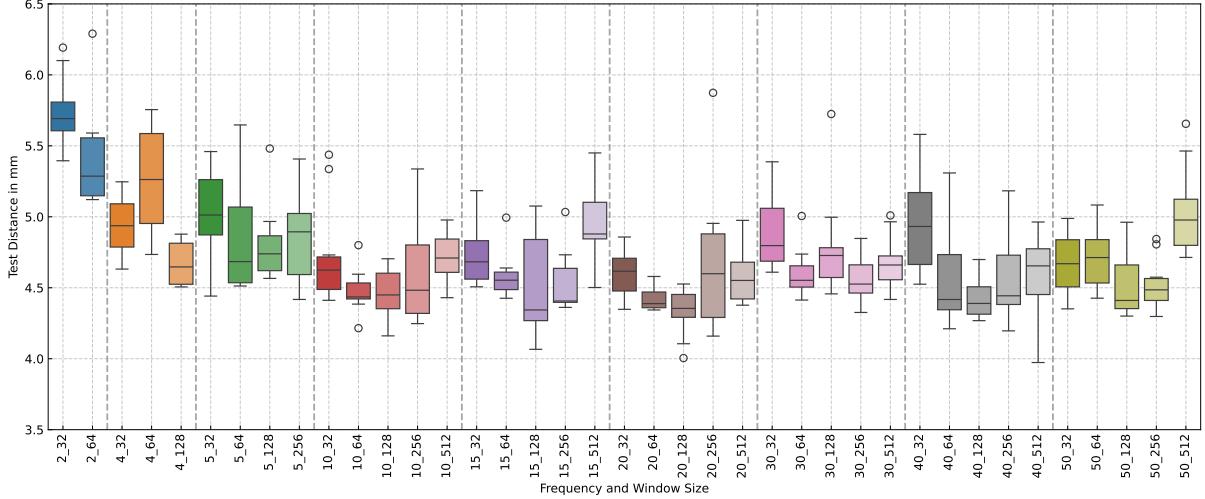
Before benchmarking the tasks’ performance, we analyzed the spectral characteristics of the raw data to determine the optimal input frequency and STFT configuration. We first examined the frequency content of the raw 50 kHz signals. As illustrated in Figure 3, frequencies above 20 kHz exhibit negligible signal magnitude (below  $-40$  dB).



**Fig. 3** Magnitude spectra (in dB) of all seven channels. Each spectrum is normalized such that the maximum magnitude is 0 dB. Frequency is shown in kHz.

To evaluate the impact of signal processing choices, we trained multiple neural networks (10 random seeds per configuration) while varying both the input frequency and the STFT window size ( $n_{\text{fft}}$ ). For this parameter sweep, we utilized a fixed validation protocol: the ‘soft plastic’ interactions were held out as the test set, and the networks were trained on the remaining materials. Figure 4 reports the Euclidean distance (in mm) averaged across this held-out set over all 10 seeds.

The results indicate that using only low-frequency components (below 10 kHz) degrades localization accuracy. However, incorporating higher-frequency components beyond 20 kHz does not lead to any measurable performance improvements. Based on these findings, we selected 20



**Fig. 4** Test distance by frequency and window size ( $n_{\text{fft}}$ ). Mean distances (mm) with 10 repetitions, showing the influence of frequency and spectral resolution.

kHz for our subsequent analyses. This choice balances accuracy with computational efficiency by excluding frequencies that provide little or no useful information. Similarly, we chose a window size equal to 128, as it offered a favorable trade-off between time-frequency resolution and model performance in our preliminary evaluations, achieving an average Euclidean distance error of 4.332 mm.

## 4.2 Impulse Response Localization Task

To evaluate the model’s ability to generalize across different surface properties, we extended our experimental analyses beyond the initial soft plastic evaluation scenario used for parameter tuning. We adopted a “leave-one-material-out” strategy to test unseen material categories. Specifically, we conducted separate experiments in which soft plastic, hard plastic, wood, and metal were individually held out as the test set, while the network was trained on the remaining materials. This approach rigorously tests the model’s robustness to varying vibro-acoustic impedances and surface textures. Each configuration was trained using 10 different random seeds to ensure statistical reliability; the mean localization errors and standard deviations are summarized in Table 1.

As presented in Table 1, our findings reveal a strong correlation between the model’s localization accuracy and the physical properties of the

**Table 1** Results on impulse response localization dataset: mean squared error (MSE) and Euclidean distance in mm across different test splits. The results are calculated over 10 different seeds. Lower MSE values indicate better performance. The values in parentheses represent the standard deviation.

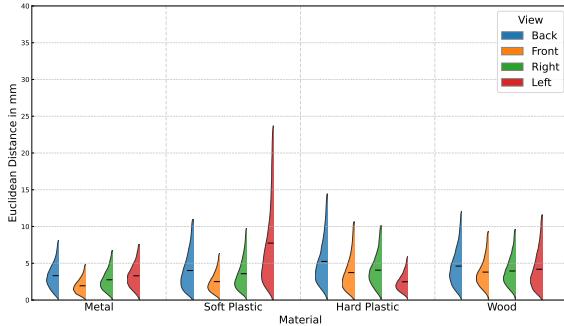
| Test Split   | MSE ↓                | Euclid. Dist. ↓      |
|--------------|----------------------|----------------------|
| Metal        | <b>0.007 (0.011)</b> | <b>3.460 (0.681)</b> |
| Soft Plastic | 0.012 (0.010)        | 4.943 (0.799)        |
| Hard Plastic | 0.020 (0.007)        | 5.391 (0.785)        |
| Wood         | 0.022 (0.011)        | 5.823 (1.361)        |

indenter material. This trend underscores that the model’s performance is fundamentally tied to the distinct acoustic signature generated by each material’s interaction with the surface. The lowest error (3.460 mm) was achieved with the metal indenter. Physically, the high stiffness and hardness of metals result in a sharp, high-energy impact that propagates efficiently, producing a clean and highly localized vibrational signal with minimal ambiguity.

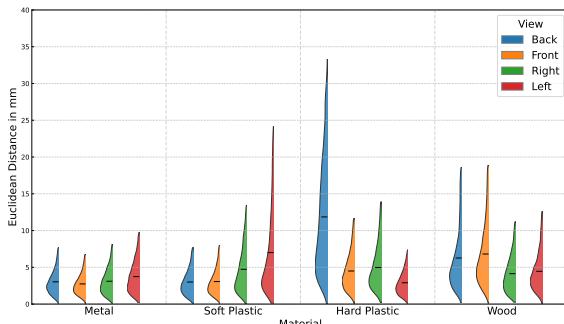
Conversely, the wood indenter yielded the highest mean error (5.823 mm), followed by hard plastic (5.391 mm) and soft plastic (4.943 mm). This suggests that materials with lower density or acoustic impedance closer to that of the robotic shell itself

may generate signals that are harder to distinguish from structural reverberations.

To further investigate these results, we analyzed the per-view prediction error (Euclidean distance in mm) for the forearm (Fig. 5) and the hand (Fig. 6). This visualization breaks down the error further by the ‘Back’, ‘Front’, ‘Right’, and ‘Left’ views.



**Fig. 5** Forearm localization error: half violin plots showing the distribution of prediction error (Euclidean Distance in mm) for four materials on the forearm section, broken down by view.



**Fig. 6** Hand localization error: half violin plots showing the distribution of prediction error (Euclidean Distance in mm) for four materials on the hand section.

The data indicates that the metal indenter provides the most spatially consistent acoustic signature for the impulse response localization task. As observed across all plots, metal achieves the lowest mean error and, notably, the lowest variance across all four views. The error distributions are tightly clustered near a low value (typically under 5 mm), suggesting that the high stiffness of

the indenter minimizes deformation upon contact. This results in a sharp, high-energy signal transfer that remains consistent regardless of the contact location.

In comparison, the polymer and organic materials exhibit distinct view-dependent distributional characteristics:

- **Soft plastic:** The distribution for soft plastic highlights the influence of material compliance. We observe increased variance, specifically in the ‘Left’ view, particularly in the Forearm plot. Physically, the softer material deforms upon impact, resulting in a damped interaction with less distinct transient features than with stiffer materials like metal. This lack of sharpness appears to make the signal more difficult to localize precisely, particularly on the left side of the robot, where structural geometry may further complicate signal propagation.
- **Hard plastic:** While this material yields precise localization in most views, the analysis reveals a specific interaction dynamic in the ‘Back’ view. As shown in the Hand (Fig. 6) and Forearm (Fig. 5) configurations, the error distribution for this view extends noticeably (reaching 30–35 mm). This suggests that the acoustic impedance match between the hard plastic indenter and the robot’s dorsal shell generates a signal propagation pattern that is more complex to spatially resolve than interactions on the ventral or lateral surfaces.
- **Wood:** Finally, the wood indenter shows higher overall error but smaller distributional tails and less extreme outliers than soft plastic ‘Left’ view and the hard plastic ‘Back’ view. Compared to metal, it exhibits higher overall variance, particularly in the ‘Front’ and ‘Back’ views of the forearm. This characteristic variance aligns with the material’s structural properties. Unlike metal and Plastic, which are manufactured to be homogeneous and isotropic, wood is a naturally heterogeneous material with variable density and stiffness across its geometry. These inherent irregularities introduce slight inconsistencies in the impact dynamics, resulting in a less deterministic acoustic signature than that of the uniform metal indenter.

### 4.3 Trajectory Tracking Task

For the trajectory tracking task, we evaluated our model using a similar ‘leave-one-material-out’ approach. We conducted four separate experiments, each time holding out one indenter material (soft plastic, hard plastic, wood, or metal) as the test set while training on the remaining three. We used 276 of the 345 categories of the Quick Draw dataset Jongejan et al. (2016) for training, 35 for validation, and a distinct set of 34 for testing.

Furthermore, to assess the model’s robustness against the hand’s own motion, we evaluated all test splits under three conditions. The first was a stationary scenario, in which the hand was turned on and remained in a fixed position. The second was a dynamic scenario, with the hand moving to randomized poses between interactions. The final, mixed scenario, utilized a combined dataset from both conditions to evaluate generalization across different motion contexts.

We report the quantitative results in Table 2, including the mean squared error (MSE) loss and the Euclidean distance (in mm) calculated over the test splits for 10 different seeds.

A critical finding in this experiment is the reversal of material performance rankings compared to the impulse response localization task, especially in the first test scenario with fixed position. While metal was the superior indenter for stationary poking, wood outperforms all other materials in this task, achieving the lowest error in every scenario (2.226 mm in ‘Fixed Position’), with soft/hard plastics performing in between. This shift highlights the fundamental physical difference between the two tasks. The impulse response localization task relies on impact dynamics, where metal’s stiffness produces a sharp, high-bandwidth impulse. In contrast, the trajectory tracking task is driven by friction and surface interaction. The natural surface roughness and grain of the wood indenter generate a rich, continuous acoustic texture as it drags across the robot’s surface, analogous to how biomimetic fingerprint ridges amplify structure-borne vibrations during sliding interactions Juiña Quilachamín and Navarro-Guerrero (2023), providing dense spectro-temporal features for tracking. Metal, being smoother, generates less distinct friction-induced vibrations when sliding, making the continuous path harder to reconstruct.

However, in the ‘Random Movement’ and ‘Fixed + Movement’ scenarios, hard and soft plastics show slightly higher errors than metal (e.g., Soft plastic: 12.946mm vs. Metal: 10.818mm in ‘Random Movement’). This may arise from excessive noise caused by minor, inconsistent hand movements across test set collections, where slight variations in individual hand poses can persist despite efforts to collect identical interactions.

Crucially, despite these challenges, the system maintains effective localization accuracy. Even in the extreme ‘Random Movement’ condition, errors remain reasonably low (e.g., 9.911 mm for wood), validating that vibrational signal analysis is a robust sensing modality capable of providing robots with reliable physical awareness even during active operation.

To contextualize these metrics, Figure 7 visualizes trajectory reconstructions for four representative classes: Baseball, Zigzag, Banana, and Keyboard.

In the ‘Fixed Position’ results, the model demonstrates high fidelity. The predicted path tightly matches the target, accurately capturing the continuous curvature of the Banana and Baseball as well as the sharp, distinct corners of the Keyboard. In the ‘Random Movement’ scenario, the model retains the ability to track the underlying trajectory despite kinematic noise. Although the predictions exhibit increased variance and local deviations due to acoustic interference, the global topology of shapes like the Zigzag remains recognizable. The system recovers the overall structure of the path, identifying that the degradation is primarily manifested as jitter rather than a total loss of spatial coherence.

Based on these results, localization accuracy is governed by the interplay between the robot’s kinematic state and the contact material properties. Across all material types, the ‘Fixed Position’ experiment consistently yielded the highest precision. The ‘Random Movement’ scenario, considerably more extreme than typical manipulation tasks, produced the highest error rates due to acoustic interference from internal motor actuation and structural vibrations. The ‘Fixed + Movement’ scenario yielded intermediate performance, suggesting that the model can learn to partially generalize across varying noise conditions.

**Table 2** Results on trajectory tracking dataset: mean squared error (MSE) and Euclidean distance in mm across different test splits and scenarios (10 seeds).

| Test Split   | Scenario         | MSE ↓                | Euclid. Dist. ↓      |
|--------------|------------------|----------------------|----------------------|
| Metal        | Fixed Position   | 0.015 (0.002)        | 3.696 (0.191)        |
|              | Random Movement  | 0.130 (0.027)        | 10.818 (1.199)       |
|              | Fixed + Movement | 0.048 (0.004)        | 5.639 (0.206)        |
| Soft Plastic | Fixed Position   | 0.013 (0.001)        | 3.472 (0.067)        |
|              | Random Movement  | 0.186 (0.018)        | 12.946 (0.719)       |
|              | Fixed + Movement | 0.085 (0.003)        | 7.297 (0.169)        |
| Hard Plastic | Fixed Position   | 0.008 (0.001)        | 2.692 (0.118)        |
|              | Random Movement  | 0.177 (0.011)        | 12.418 (0.392)       |
|              | Fixed + Movement | 0.080 (0.007)        | 6.752 (0.232)        |
| Wood         | Fixed Position   | <b>0.005 (0.001)</b> | <b>2.226 (0.079)</b> |
|              | Random Movement  | <b>0.117 (0.015)</b> | <b>9.911 (0.723)</b> |
|              | Fixed + Movement | <b>0.042 (0.002)</b> | <b>4.830 (0.088)</b> |

## 5 Conclusion

This paper demonstrates that high-accuracy touch localization on a robotic hand is achievable through vibrational signal analysis, offering a scalable, cost-effective alternative to complex tactile skin arrays. By leveraging an Audio Spectrogram Transformer (AST) to process vibrational signals from simple piezoelectric microphones, our method achieves robust performance across varied interaction modalities.

A key contribution of this work is the comprehensive analysis of how material properties and physical interaction modes influence vibro-acoustic sensing. Our results reveal a fundamental distinction between stationary and dynamic sensing: stiff materials (such as metal) generate the sharpest impulse responses for impulse-response localization, whereas textured materials (such as wood) produce the most distinct friction-based features for trajectory tracking. These findings highlight that acoustic sensing captures rich physical data beyond simple coordinates, intrinsically encoding the mechanical properties of the contact object.

We further validated the system’s robustness in active scenarios. While the introduction of motion and internal motor noise inevitably degrades precision, the system maintains effective localization accuracy (typically under 12 mm even under extreme, arguably unrealistic conditions), proving its viability for real-world tasks where robots must sense while grasping or manipulating objects. Our signal processing analysis confirms that a sampling rate of 20 kHz is sufficient to capture these

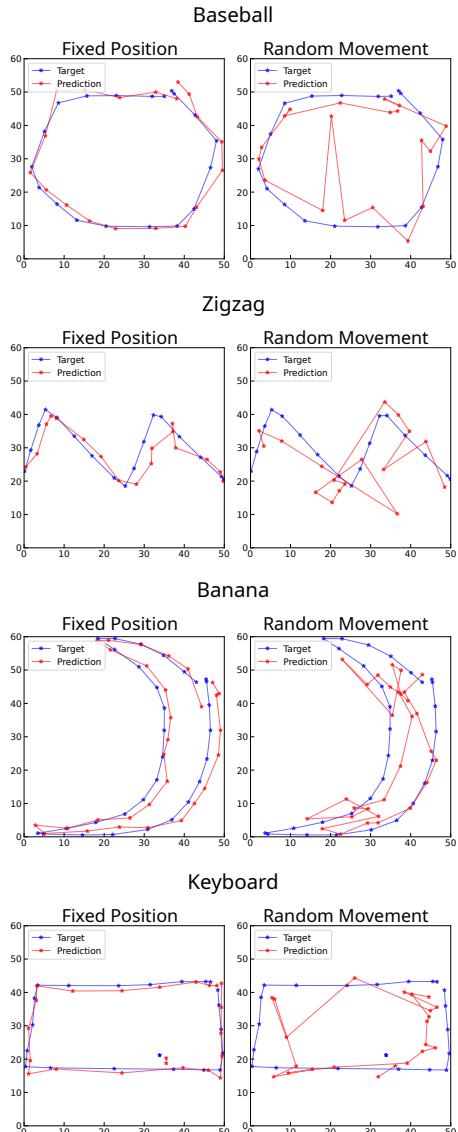
features, balancing computational efficiency and sensing fidelity.

Despite these findings, limitations remain. Our geometric analysis identified distinct acoustic interaction behaviors, including the impedance mismatch observed with hard plastic on the dorsal shell, and the signal damping caused by soft materials in complex structural areas. Additionally, although the system demonstrates resilience to self-generated motor noise, the localization error inevitably increases during motion. Future work will address these challenges by employing adaptive pre-filtering to decouple internal motor noise and fusing vibro-acoustic features with visual data to resolve geometric ambiguities. Furthermore, we aim to extend the system’s capabilities to include closed-loop manipulation and slip detection, thereby enhancing robustness during delicate and clutter-rich interaction tasks.

Finally, to facilitate reproducibility, we are making our model checkpoints, datasets, and experimental setups publicly available. By open-sourcing these resources, we aim to accelerate the development of affordable whole-body contact perception, which is essential for advancing robotic grasping and manipulation capabilities.

## Declarations

**Funding.** This research was partially funded by the Bundesministerium Forschung, Technologie und Raumfahrt (BMFTR) under the Programme DATIPilot Innovationsprints project No. 03DPS1242A (Vibro-Sense).



**Fig. 7** Comparison of target (blue) vs. predicted (red) trajectories for fixed and random movement scenarios across four illustrative examples: Baseball, Zigzag, Banana, and Keyboard.

**Conflict of interest/Competing interests.** The authors declare that they have no conflict of interest.

**Ethics approval.** This article does not contain any studies with human participants or animals performed by any of the authors.

**Open Access.** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or

format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third-party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

**Consent to participate.** Not applicable

**Consent for publication/Informed consent.** Not applicable

**Availability of data and materials.** Full dataset will be provided for the camera-ready version of this manuscript.

**Code availability.** Full code and checkpoints will be provided for the camera-ready version of this manuscript.

## References

Abraira, V.E., Ginty, D.D.: The Sensory Neurons of Touch. *Neuron* **79**(4), 618–639 (2013)

Bonner, L.E.R., Buhl, D.D., Kristensen, K., Navarro-Guerrero, N.: AU Dataset for Visuo-Haptic Object Recognition for Robots. Dataset Description arXiv:2112.13761, Aarhus University, Denmark (2021). <https://doi.org/10.48550/arXiv.2112.13761>

Chen, Y., Sun, Y., Wei, Y., Qiu, J.: How Far for the Electronic Skin: From Multifunctional Material to Advanced Applications. *Advanced Materials Technologies* **8**(8), 2201352 (2023)

Gong, Y., Chung, Y.-A., Glass, J.: AST: Audio Spectrogram Transformer. In: Interspeech, Brno, Czech Republic, pp. 571–575 (2021). <https://doi.org/10.21437/Interspeech.2021-698>

Hoffmann, M., Longo, M.R.: Body Models in Humans and Robots. In: The Routledge Handbook of Bodily Awareness, 1st edn., p. 13.

Routledge, London (2022). <https://doi.org/10.4324/9780429321542>

Hardman, D., Thuruthel, T.G., Iida, F.: Multimodal Information Structuring with Single-Layer Soft Skins and High-Density Electrical Impedance Tomography. *Science Robotics* **10**(103), 2303 (2025) <https://doi.org/10.1126/scirobotics.adq2303>

Juña Quilachamín, O.A., Navarro-Guerrero, N.: A Biomimetic Fingerprint for Robotic Tactile Sensing. In: International Symposium on Robotics (ISR Europe), pp. 112–118. VDE Verlag GmbH, Stuttgart, Germany (2023). <https://doi.org/10.48550/arXiv.2307.00937>

Jongejan, J., Rowley, H., Kawashima, T., Kim, J., Fox-Gieg, N.: The Quick, Draw! - A.I. Experiment. <https://quickdraw.withgoogle.com/>. Accessed: 2025-09-12 (2016)

Lu, S., Culbertson, H.: Active Acoustic Sensing for Robot Manipulation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3161–3168 (2023)

Liu, J., Chen, B.: SonicSense: Object Perception from In-Hand Acoustic Vibration (2024)

Lu, S., Chen, Y., Culbertson, H.: Towards Multisensory Perception: Modeling and Rendering Sounds of Tool-Surface Interactions. *IEEE Transactions on Haptics* **13**(1), 94–101 (2020)

Lambeta, M., Chou, P.-W., Tian, S., Yang, B., Maloon, B., Most, V.R., Stroud, D., Santos, R., Byagowi, A., Kammerer, G., Jayaraman, D., Calandra, R.: DIGIT: A Novel Design for a Low-Cost Compact High-Resolution Tactile Sensor with Application to In-Hand Manipulation. *IEEE Robotics and Automation Letters* **5**(3), 3838–3845 (2020)

Lee, M., Yoo, U., Oh, J., Ichnowski, J., Kantor, G., Kroemer, O.: SonicBoom: Contact Localization Using Array of Microphones. *IEEE Robotics and Automation Letters* **10**(7), 7603–7610 (2025) <https://doi.org/10.1109/LRA.2025.3576067>

Mejia, J., Dean, V., Hellebrekers, T., Gupta, A.: Hearing Touch: Audio-Visual Pretraining for Contact-Rich Manipulation (2024)

Navarro-Guerrero, N., Toprak, S., Josifovski, J., Jamone, L.: Visuo-Haptic Object Perception for Robots: An Overview. *Autonomous Robots* **47**(4), 377–403 (2023) <https://doi.org/10.1007/s10514-023-10091-y>

Google <https://developers.google.com/optimization>

Pop, P.C., Cosma, O., Sabo, C., Sitar, C.P.: A Comprehensive Survey on the Generalized Traveling Salesman Problem. *European Journal of Operational Research* **314**(3), 819–835 (2024) <https://doi.org/10.1016/j.ejor.2023.07.022>

Seed Robotics, S.: RH8D Adult Size Dexterous Robot Hand. <https://www.seedrobotics.com/rh8d-adult-robot-hand>

Toprak, S., Navarro-Guerrero, N., Wermter, S.: Evaluating Integration Strategies for Visuo-Haptic Object Recognition. *Cognitive Computation* **10**(3), 408–425 (2018) <https://doi.org/10.1007/s12559-017-9536-7>

Touchlab. <https://www.touchlab.io/>. Accessed: 2025-09-22

Universal Robots, U.: UR5e Lightweight, Versatile Cobot. <https://www.universal-robots.com/products/ur5e/>

Wall, V., Brock, O.: A Virtual 2D Tactile Array for Soft Actuators Using Acoustic Sensing. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10029–10034 (2022)

Wang, A., Gollakota, S.: MilliSonic: Pushing the Limits of Acoustic Motion Tracking. In: CHI Conference on Human Factors in Computing Systems. CHI '19, pp. 1–11. Association for Computing Machinery, Glasgow, Scotland, UK (2019). <https://doi.org/10.1145/3290605.3300248>

Wang, L., Ma, L., Yang, J., Wu, J.: Human Somatosensory Processing and Artificial Somatosensation. *Cyborg and Bionic Systems*

**2021** (2021)

Wall, V., Zöller, G., Brock, O.: Passive and Active Acoustic Sensing for Soft Pneumatic Actuators. The International Journal of Robotics Research **42**(3), 108–122 (2023) <https://doi.org/10.1177/02783649231168954>

Yuan, W., Dong, S., Adelson, E.H.: GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force. Sensors **17**(12), 2762 (2017)