SoundScroll: Robust Finger Slide Detection Using Friction Sound and Wrist-Worn Microphones

Daehwa Kim*
Carnegie Mellon University
Pittsburgh, PA, USA
daehwak@cs.cmu.edu

Eric Whitmire
Meta Reality Labs
Redmond, WA, USA
ewhitmire@meta.com

Roger Boldu
Meta Reality Labs
Redmond, WA, USA
rboldu@meta.com

Wolf Kienzle
Meta Reality Labs
Redmond, WA, USA
wkienzle@meta.com

Hrvoje Benko
Meta Reality Labs
Redmond, WA, USA
benko@meta.com

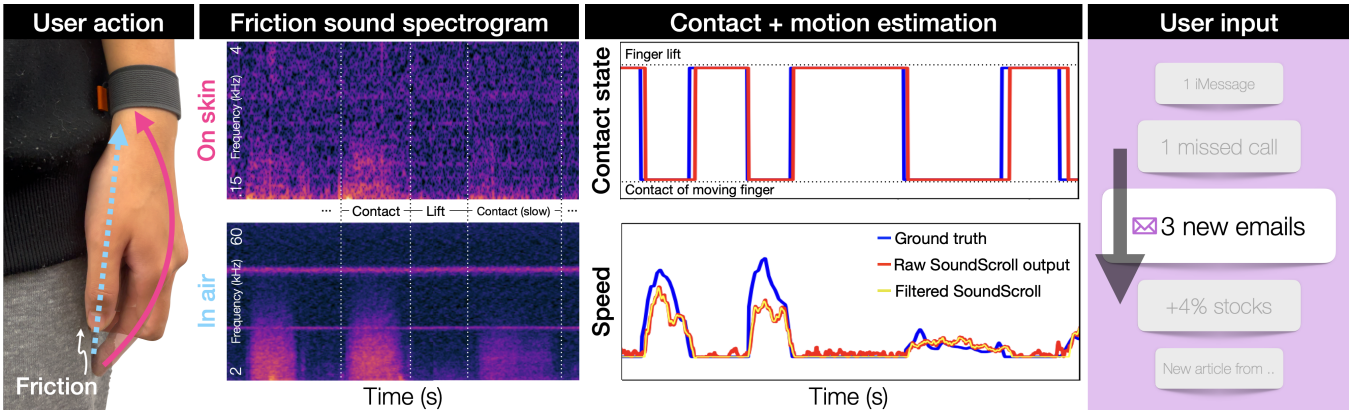

Figure 1: A conceptual illustration of SoundScroll system. (User action) A sliding finger on the surface creates the vibration induced by friction. The vibration propagates on the skin and in the air. (Friction sound spectrogram) Wrist-worn acoustic sensors capture the vibration. (Contact + motion estimation) With dual-channel audio measurements, our system estimates the moving finger's contact state and sliding speed. (User input) Finally, users can control user interfaces such as scrolling a list.

ABSTRACT

Smartwatches have firmly established themselves as a popular wearable form factor. The potential expansion of their interaction space to nearby surfaces offers a promising avenue for enhancing input accuracy and usability beyond the confines of a small screen. However, a key challenge is in detecting continuous contact states with the surface to inform the start and end of stateful interactions. In this paper, we introduce SoundScroll, enabling a rapid and precise determination of contact state and fingertip speed of sliding finger. We leverage vibrations from friction between a moving finger and a surface. Our proof-of-concept wristband captures a dual-channel vibration signal for robust sensing, considering both on-skin and in-air components. Our software predicts a finger sliding state as fast as 20 ms with an accuracy of 93.3%. Augmenting prior approaches detecting tap events, SoundScroll can be a robust, low-latency, and precise contact and motion sensing technique.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing systems and tools.*

KEYWORDS

Contact; Surface input; Slide; Indirect input; Wristband;

ACM Reference Format:
Daehwa Kim, Eric Whitmire, Roger Boldu, Wolf Kienzle, and Hrvoje Benko. 2024. SoundScroll: Robust Finger Slide Detection Using Friction Sound and Wrist-Worn Microphones. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers (ISWC '24), October 5–9, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3675095.3676614

*This work was done during an internship at Meta Reality Labs.

ISWC '24, October 5–9, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1059-9/24/10.
https://doi.org/10.1145/3675095.3676614

1 INTRODUCTION

Smartwatches have become a popular form factor and are ubiquitous in the consumer market. Despite their inherent advantages in portability and wearability, their input space is confined to a small screen. Over the last decade, researchers have expanded the interactable surface area to near-hand surfaces [16, 34, 40, 59] enabling

touchpad-like interactions. This has ergonomic advantages [3, 26], helps precise selection with closed-loop touch feedback [23], and provides contextually adaptive surfaces [41]. A plethora of sensing approaches have been explored to enable this, ranging from camera-based solutions [21, 32, 45, 54, 56], motion sensor [16, 34, 40, 59], electric [17], to electromagnetic field [2, 35, 52] sensing.

Despite significant advancements, accurately identifying the duration of an interaction without instrumenting the environment remains challenging. This identification is crucial for enabling stateful touch-based input, which facilitates a range of dynamic interactions such as list scrolling, drag-and-drop, and two-dimensional pointing [1]. These interactions are essential for effective user interfaces in modern applications. Accurate detection requires knowing whether a fingertip is currently interacting with the surface (e.g., scrolling, dragging) or has stopped. Previous wristband technologies have primarily been limited to sensing tap events; however, reliably identifying the end of interaction (stopping scroll or touch-up events) has remained a considerable challenge [11]. Inaccurate or delayed detection of touch events can lead to a phenomenon known as the "serif effect" [37]. In the context of dragging interactions, this effect can cause unintended backscrolling during repeated scrolling gestures, leading to a frustrating user experience. This underscores the need for a reliable, low-latency sensing source that detects the continuous contact state of a finger to ensure a complete and compelling user experience.

In response, we propose SoundScroll, a wristband solution to continuously detect contact states of sliding fingers for supporting stateful touch interaction. SoundScroll relies on the mechanical energy produced when a finger slides over a surface to detect contact and motion. This energy induces vibrations through the skin and air, and SoundScroll captures these vibrations with two microphones—skin-contact and in-air microphones. Note, as this vibration is key in SoundScroll, it works while a finger is in motion. We designed a proof-of-concept wristband prototype and verified its accuracy across several common surface materials and two warm postures along with various users and speed conditions. We designed a proof-of-concept wristband prototype and verified its accuracy across several common surface materials and two warm postures along with various users and speed conditions. We designed a proof-of-concept wristband prototype and verified its accuracy across several common surface materials and two warm postures along with various users and speed conditions. We summarize the main contributions of this paper:

(1) A wristband sensing solution that continuously detects the sliding finger's contact state to inform the start and end of stateful user inputs, especially for scrolling interaction.
(2) A fast estimation of finger contact state in as little as 20 ms.
(3) A multi-task model that classifies finger contact while simultaneously regressing an instantaneous fingertip speed.
(4) The evaluations that show the resilience of SoundScroll to four surface types and two common arm postures compared to a kinematics approach.

In the rest of the paper, we review relevant usability issues and technical approaches. Then, we describe the principle of friction-based sensing. Subsequently, we present the SoundScroll prototype and the results of the associated user studies. Finally, we discuss the future opportunity to integrate our technique into existing technologies.

2 RELATED WORK

2.1 Non-Wristband Solutions

While a computer vision method with a head-mounted camera is accurate for tracking the finger's 2-dimensional position, it often struggles to distinguish whether the finger is touching a surface or hovering just a few millimeters above it [12, 54]. Recent work embedded an active illuminant to the wristband [21, 46] to improve surface contact detection. However, occlusion and field-of-view limitations persist as challenges. Other popular approaches instrument parts of the user's finger directly, including the fingerpad [31, 56], fingernail [11, 40], or a ring [16, 17, 22, 34, 35, 49, 51]. Among them, an electrical method [17, 51, 58, 58] detects contact loops for contact sensing but is limited to the body or surfaces with specific electrical properties. One can also use a ring with a motion sensor, such as an inertial measurement unit (IMU). This method adopts a kinematics approach, leveraging the finger joints' physical structure to gather crucial information [16, 39]. The methods, however, rely on finger-instrumented devices, which can encumber and restrict the dexterity of the hand for manual activities [42] or reduce tactile sensitivity.

2.2 Wristband Solutions

Smartwatches have been successfully established as a friendly form factor in the consumer market. Sensing the finger using sensors embedded in the wristband becomes more desirable for users [55]. Although finger-joint tracking with wrist-worn sensors has been demonstrated [5, 14, 18, 19, 24, 53], knowing the finger's contact condition with the world remains a separate challenge. Optical methods, such as employing an IR light and a camera positioned under the wrist [32, 38, 50], offer a means to detect finger contact by identifying illumination changes. However, this method is sensitive to curved or uneven surfaces, varying illumination conditions, and the camera's angle of attack, making this less suitable for dynamic mobile environments. In contrast, acoustic and motion sensors have well shown their promise as units to inform touch for the wristband [8, 13, 26, 29, 32, 44]. However, the majority of wristband solutions beyond on-skin interactions have focused on detecting discrete events, particularly simple taps. Skinput [29] uses an armband and bio-coustics transmission to enable on-skin tap interaction. TapID [26] capture bio-acoustic signal created when a user tap different fingers on the surface and identify a tapping finger.n Acustico [8] implemented a wrist-worn device that can resting on the table during interaction and enables tap and swipe detection. Similarly, Anywhere Surface Touch [32] detects left and right swipes and taps acoustic waves propagating through a desk along with camera data. In contrast, SoundScroll has several distinct advantages. Firstly, SoundScroll utilizes vibrations propagating through skin and air, which allows for use without requiring the user's wrist to remain stationary on a desk, thus supporting broader mobile scenarios. Secondly, SoundScroll employs low-profile, non-camera-based sensors, making it more suitable for integration into wearable devices. Lastly, while Anywhere Surface Touch introduces a delay in discrete gesture recognition (i.e., feedback comes after

the gesture ends) that can negatively affect performance in tasks such as scrolling, SoundScroll continuously updates finger contact state and speed every 20 ms, enabling more fluid and interactive control. AO-Finger [55] also noted the significance of fine-grained UI control by enabling continuous thumb swipe tracking using a single hand. However, its focus is primarily on thumb-to-finger touch interactions, without considering interactions with external surfaces. Moreover, AO-Finger relies solely on an on-skin microphone (stethoscope) for detecting finger contact. They noted that motion artifacts, such as tendon movements when a finger moves without touching, affect the accuracy of contact recognition. In contrast, SoundScroll incorporates an additional in-air microphone, which potentially helps address motion artifacts related to finger movements that occur without actual contact. Furthermore, SoundScroll not only detects contact but also predicts fingertip motion speed with a single model, enhancing interactions with features like inertial scrolling.

3 METHODS

When a finger slides over a surface, vibration is induced by the friction interaction as the finger traces grooves on textures. This vibration will propagate through the skin, the touched surface, and air. While capturing vibration by instrumenting the surface is effective, it is not suitable for wearable applications. Thus, we leverage the other two path's vibrations captured by on-skin or in-air microphones. The past work has shown the relationship of acoustic signatures to contact state or surface properties [33, 48]. SoundScroll employs passive acoustic sensing using two wrist-worn microphones to capture this friction-induced vibration. The captured friction sound profile shows different powers, frequency ranges, and resonance depending on finger contact interactions as shown in Figure 1. We take a data-driven approach to predict a fingertip sliding state using the sensors on the wrist.

3.1 Sensor Positions

To test the impact of on-skin microphone positions, we conducted a pilot study with one of the authors. For ground-truth finger contact and sliding speed collection, we used the Sensel Morph touchpad. We placed the Sonion Voice Pick-Up (VPU) contact microphone to four candidate positions on the wrist: palmar, dorsal, radial, and ulnar sides. For the palmar side, a microphone was located around the flexor tendon, as reported in the literature for its effectiveness to capture the vibration signal propagated from a fingerpad [4, 55]. For the ulnar side, a sensor was placed on the bone to see vibration properties propagating through the bone [13].

We evaluated the performance of each on-skin microphone position with the same data collection procedures and metrics in the main study but only on one texture (; a bare touchpad) and with one of authors. In total, 480 sliding instances (4 directions × 10 trials × 2 speeds × 6 re-worn sessions containing two arm postures) for each comparison condition were collected. We compared each microphone position in a leave-one-session-out cross-validation scheme trained on ExtraTreesClassifier [20] with default parameters of Python Scikit Learn. The result shows that a VPU on the palmar side produces the highest accuracy among the four locations as shown in Figure 2 left. This result also aligns with prior literature using a stethoscope microphone [4, 55].

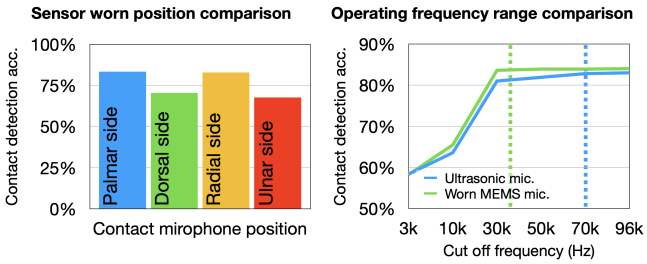

Figure 2: Results comparing contact state estimation performance on four sensor worn location (left) and frequency range (right).

3.2 Frequency Range

In this pilot study, we find the frequency range that largely contributes to contact state estimation. For the in-air microphone, we used the SPH8878LR5H-1 Analog MEMS Microphone. This microphone is sensitive to the frequency range from 7 Hz to 36 kHz. However, we could still observe the friction sound up to around 55 kHz and used the full range (up to 96 kHz) for analysis. For reference purposes, we also set up an ultrasonic microphone that can reliably capture sound up to 70 kHz. For the on-skin microphone, our VPU exceeds the necessary range (up to 10 kHz) for capturing vibrations transmitted through the human body (up to around 5 kHz [47, 57]), thus we conducted this study only for an in-air microphone.

We evaluated the performance of each on-skin microphone position with the same procedures in the Sensor Position pilot study. We conducted an ablation study with six frequency ranges, from 0 Hz to 3, 10, 20, 50, 70, and 96 kHz. Subsequently, we trained a machine learning model (ExtraTreesClassifier) using each of these low-pass filtered sound spectrograms. The result shows that frequency ranges over 30 kHz yield a similar level of performance but cutting off lower than 30 kHz makes a large accuracy drop (Figure 2 right). This indicates that SPH8878LR5H-1 with sensitivity up to 36 kHz would be sufficient for our finger contact sensing prototype.

3.3 Wristband Hardware

The SoundScroll prototype is shown in Figure 3. The Sonion Voice Pick-Up (VPU) bone sensor [43] and SPH8878LR5H-1 breakout is placed on the palmar wrist. The wristband additionally has an Adafruit 9-DOF orientation IMU (BNO08) with Sparkfun Thing Plus ESP32 Microcontroller. This is only used as a benchmark to compare IMU's kinematics-based baseline with our friction-sound approach later in our user study. This breakout provides the orientation of the sensor and linear acceleration (without gravity), along with acceleration and angular acceleration at 100 Hz of sampling rate.

Recognizing the vital importance of ensuring reliable contact between the VPU sensor and the wrist, the VPU sensor is elevated by affixing a small rubber element. The Velcro enables adjustment of the wristband to accommodate wrists of various radii and ensures optimal contact conditions.

Figure 3: A SoundScroll proof-of-concept wristband.

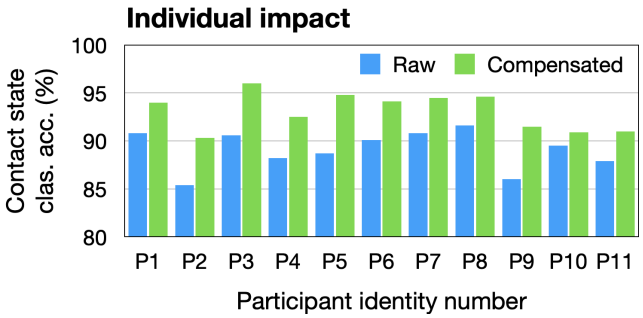

Figure 4: The sliding finger's contact state classification results per participant. The blue bar represents the classification result of the raw SoundScroll output, while the green bar is the result compensating for a 20 ms offset.

3.4 Data Processing and Deep Learning

SoundScroll captures the passive acoustics naturally induced by the user's sliding interactions on the surface. Our system employs a dual-microphone setup to reject the noise of each sensor. The on-skin microphone is sensitive to vibrations ranging from 150 Hz to 4 kHz, capturing low to mid-range frequencies directly from skin contact. Meanwhile, the in-air microphone is designed to pick up friction sounds in a higher frequency range, from 2 kHz to 60 kHz. Each audio signal is converted into the frequency domain to extract better features. We used the short-time Fourier transform (STFT) to convert the audio signals into a spectrogram. To achieve better spectrogram resolution, we use different STFT window sizes for each audio signal. For the on-skin microphone's audio, STFT takes a window size of 2048 with 384 hop length. For an in-air microphone, a smaller window size of 512 is used with the same hop length.

We use a single deep learning model to perform two tasks, estimating both finger sliding state and speed. The model uses a three-layer unidirectional Long short-term memory (LSTM) incorporating 64 hidden states and a 0.5 dropout rate. The audio representation from this LSTM is linearly transformed into a contact state classification and speed regression. The model utilizes Cross Entropy loss for contact detection and Mean Squared Error (MSE) loss for speed estimation. The summation of two losses is then used to update the weights. In training, the model takes a sound spectrogram as an input and the data undergoes augmentation through a 50-overlapping window. The batch size is 64 and weight updates are performed using the Adam optimizer with a learning rate of 0.0003. An early stopping condition is implemented, triggered when the loss improvement is smaller than 0.0001 over 10 epochs.

3.5 Post processing

We apply a filter to smooth the output contact detection. Our model output intermittently yields oscillatory predictions between touch-down and touch-up within a span of a few milliseconds. As this case rarely happens in human scrolling behavior, we have implemented a filtering mechanism, effectively smoothing out the prediction results. We take the recent 20 ms predictions to check the abnormality. If the predicted contact state is the same across 20 ms or only one state change is observed, our system immediately uses this output for interaction. If there is more than one contact-state change within 20 ms windows, our algorithm smooths this contact prediction output. The past 50 ms predictions are used to calculate the "stable" contact state and update new predictions. This smoothing algorithm maintains a prediction rate of 375 Hz when the output is stable, and it remains fast at 50 Hz even in abnormal results.

4 EVALUATION

The study aims to evaluate three aspects. We examine the capability of friction sound for (1) reliable and rapid touch sensing, (2) in different arm postures and (3) touching surfaces. We recruited 11 participants (mean age 27.4, std 2.55, min 25, max 32; two identified as female, ten as male) for the approximately 40-minute user study.

4.1 Data Collection Apparatus

For ground-truth finger touch and sliding speed collection, we used the Sensel Morph touchpad [27]. This touchpad has been used for reliable ground truth systems in prior work [9, 10, 39]. The touchpad is capable of sensing with overlaid layers [28], allowing us to test four different surface material options during data collection. Two audio input channels (on-skin and in-air) were collected over a cable with a 3.5 mm stereo headphone jack to the MOTU 1248 audio interface and sampled at 192 kHz. The acceleration, gyro, and absolute orientation values were read with a microcontroller and streamed over a USB-C cable to a laptop.

To cover diverse contact surface materials, four surfaces were used in data collection: denim, bare touchpad (polymer), paper, and cellulose tape. These materials are chosen to reflect the actual interaction scenarios that users could encounter, such as jeans and many fabrics, paper books, or glass tables or doors, etc.

The user study data was collected under two arm postures, an arm resting on the desk and floating. In each case, the sliding motion generally pivoted at either the wrist or elbow, which generates different kinematic motions for motion sensors at the wrist. This data collection condition is for comparing two prediction results from friction sound and the traditional motion sensor (in this study, IMU) approaches, respectively.

The data collection was done in an open office space with numerous people present. We did not specifically create artificial noise nor discourage any naturally generated noises. Therefore, the

dataset we used for training and evaluation involves participants' and any others' chatting, vacuuming, mechanical keyboard typing, door/drawer movements, coffee grinding, handling plastic bags, foot stepping, etc. This was intended to incorporate natural environmental noise into the study, however, there was no controlled environment in this study where noise levels were controlled.

4.2 Procedure

The participant wears a wristband and slides the finger on the touchpad to scroll through a list on the screen. The task is similar to those used in 1-dimensional Fitts' law studies, specifically involving aiming the target element in the list towards a cursor fixed in the middle of the screen. The physical finger displacement on the touchpad was linearly mapped to the visual's movement and the visuals were continuously updated. The list contains 50 elements and the target location is randomly selected among 50 elements in each selection trial. A session contains combinations of three target sizes (small, medium, large) and four scroll directions (up, down, right, and left). Four surface materials were located on the touchpad in randomized order. After completion of the task under target sizes and scroll direction combinations, a study instructor changed the surface materials on the touchpad by applying removable tape between the touchpad and the material. The participant first completed three sessions while resting their arm on the desk. After that, they completed the other three sessions while hovering their arm over the desk. To add a variation of wrist-worn location, a participant took off and re-donned the wristband after the completion of each session.

5 RESULTS

In total, participants completed 3168 scrolling tasks (11 participants × 4 surface materials × 2 arm postures × 3 target sizes × 4 scroll directions × 3 re-worn sessions). This process yields 6,034,992 data points, with each point representing 2 ms chunks. The collected data spans 3.35 hours of finger-sliding instances.

5.1 Contact State Detection

We evaluated SoundScroll's contact state detection model in a leave-one-participant-out (LOPO) cross-validation scheme. We used data from ten participants to train a model and evaluated the trained model on one left-out participant's data. This yields eleven models and we averaged the results from each. The contact state detection was evaluated every 2 ms over 3.35 hours of data, and SoundScroll achieves 89.2% of contact state detection accuracy.

We further analyze data to contextualize the result. Most of these detection errors are from a slight offset between the label and prediction, which is shown as a delay (shown in Figure 1). In drag-and-drop tasks, humans typically do not perceive delays when they are less than 33 ms for indirect input [6, 7], remains stable with direct input delays under 25 ms [15]. Thus, we set a 20 ms window as an acceptable threshold for usability. With this compensation, the contact state detection accuracy increases to 93.3% overall and exceeds 90% for each participant (Figure 4). In the rest of the result section, we describe the accuracy with 20 ms compensated.

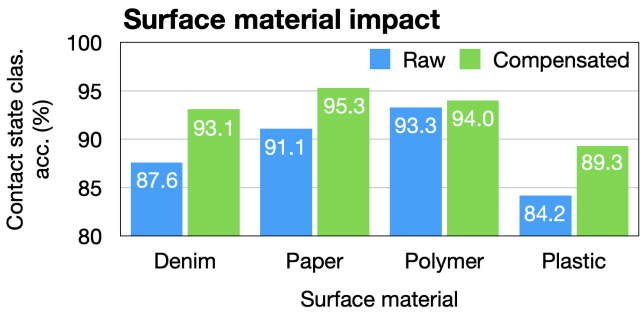

Figure 5: The sliding finger's contact state classification results per surface. The blue bar represents the classification result of the raw SoundScroll output, while the green bar is the result compensating for a 20 ms offset.

5.2 Dual-Microphone Impact

To evaluate the benefit of using a dual-microphone setup, we conducted an ablation study by training the model with each microphone's signals and analyzed its contribution. When trained only on the on-skin microphone's signal, the accuracy for detecting contact was 85.1%. Training with solely the in-air microphone resulted in a contact state detection accuracy of 92.4%. However, employing both microphones simultaneously improved the contact state detection accuracy to 93.3%. This result shows dual microphone configuration shows an improvement compared to the sole on-skin microphone configuration.

5.3 Surface Material Impact

We evaluated our approach on four surface materials, denim, paper, polymer, and plastic. The summarized result is shown in the Figure 5. We conducted our analysis using the leave-one-material-out method. When tested on denim as the surface material, we achieved a contact detection accuracy of 93.1%. For paper, the contact detection accuracy is 95.3%, which is the highest accuracy among the four surfaces. When testing on polymer, the touch detection accuracy reaches 94.0%. Finally, on surface material plastic, we attained a contact detection accuracy of 89.3%. The plastic surface yields the lowest accuracy as the surface makes a faint friction sound.

5.4 Arm Posture Impact

We analyzed results with the leave-one-posture-out method on all participant data. This led a data split of half the dataset for training and testing. Figure 6 illustrates the varying impact of arm posture on the contact state detection accuracy. Even though the training data does not include the data from other arm postures, SoundScroll shows equally accurate contact state detection accuracy for arm pivoting (93.0%) and wrist pivoting (93.5%).

5.5 Comparison to Kinematics-based Approach

For this specific analysis, we additionally trained a model under the IMU approach and compared its performance to SoundScroll's friction sound-based approach. The same network pipeline is used, yet the only difference is the input size, which is a combination of IMU's orientation, acceleration, angular acceleration, and linear

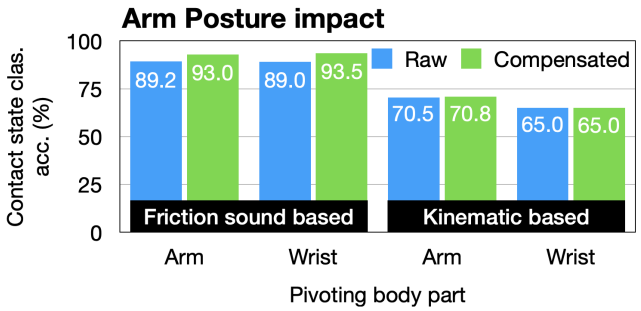

Figure 6: The sliding finger's contact state classification results comparing friction-sound (SoundScroll) and motion-based approaches. Each model was trained under two different arm postures. The blue bar represents the classification result of the raw SoundScroll output, while the green bar is the result compensating for a 20 ms offset.

acceleration. The IMU data is interpolated to 375 Hz to match the SoundScroll's prediction rate. As shown in Figure 6, the contact state detection accuracy is around 20% lower than the friction based approach.

6 DISCUSSION

Fingertip speed prediction. Knowing the fingertip's speed in scrolling is useful for richer interaction, particularly in applications like inertial scrolling. Thus, we conducted an additional analysis of how the finger vibration might provide relevant information. The same study data introduced in Section 4 was used in the training and testing of speed estimation. The speed label was calculated from the position data collected from the same ground-truth touchpad hardware. Similar to finger contact state detection evaluation, we used the LOPO scheme to evaluate speed estimation accuracy. Note, here we designed our model to be multi-task which estimates finger contact state and speed simultaneously. The mean absolute error (MAE) is 47.20 mm/s (SD=86.87), while the label speed is 96.7 mm/s on average (SD=169.6). SoundScroll achieved fairly accurate fingertip speed estimation and an example speed estimation result is illustrated in Figure 1.

Friction sound approach v.s. Kinematics approach. One common method for surface input is the kinematic approach, which involves installing motion sensors (e.g., IMUs) on the finger. This technique leverages the constraints of hand ergonomics to estimate the finger's state. However, this approach may not be suitable for wristband form factors, as resting the wrist on a desk makes it challenging to capture motions accurately. SoundScroll leverages the nature of the sound-based approach, where the audio signals are directly influenced by the interaction between the fingerpad and the surface, rather than by ergonomics. The friction-sound approach could potentially reduce the required amount of training data compared to IMU approaches, as it does not necessitate data collection in various arm posture conditions.

System Delay. Our system takes a new spectrogram input for every incoming set of 512 audio samples. This enables us to achieve 375 Hz (= 192000 Hz sampling rate / 512 window size; account for 2.7 ms) of touch and speed prediction rate. The preprocessing (STFT) takes 0.2 ms for the corresponding window size of sound signals. The inference time of our model in Apple Core ML is 2.0 ms on an Apple Macbook Pro with an M1 processor. These are roughly faster than the input stream time of 2.7 ms. The input device's system delay affects human perception (~5-10 ms for direct input [30]; ~33 ms for indirect input [6, 7]) and selection performance (25 ms for direct input [15]; 75-100 ms for indirect input [25, 36]). SoundScroll's update rate can be as low as 2.7 ms, which is advantageous in mitigating the delay effect when applying smoothing filters.

SoundScroll and scrolling. SoundScroll is especially useful when the interaction includes repeated motions in different directions. For example, when users scroll a long list with clutching, they perform swipes in one direction to scroll and move the finger back to the original position in the opposite direction and repeat. The delayed or inaccurate detection of the contact state creates unexpected list rewinding at the end of the scroll as the finger returning is captured as an unfinished input. SoundScroll's fast prediction can resolve this usability issue.

Toward Full Functionality. SoundScroll can be combined with other modalities to achieve a general-purpose 2-dimensional input device. Two more modules would be desired — detecting taps and tracking finger motion direction. The primary limitation of SoundScroll is that it detects the contact state of a finger in motion. SoundScoll itself would be challenging to support certain user inputs such as long-press. We envision combining the tap-sensing module with SoundScroll so that the system enhances robustness and provides full functionalities. Several earlier works [26, 44] already demonstrated tap interaction by embedding IMU or other acoustic sensors. Those approaches can also be integrated with SoundScroll. Detecting the finger sliding direction from the wrist could be more challenging. A combination of IMU, infrared, optical flow, or depth sensors can be employed to estimate the distance change of the back of the hand from the sensor and determine the moving directions. Based on studies like Back-Hand-Pose [53], which can estimate full hand pose by observing only a part of the back of the hand, it may be feasible to determine the direction of the moving finger. Nonetheless, SoundScroll simplifies the requirements to classifying movement directions instead of estimating both precise displacement and direction, making the challenge more manageable. Lastly, our studies were focused on scrolling interactions. While the principle may remain the same to detect a moving finger's contact state in various applications like drag-and-drop, two-dimensional pointing, and drawing, further thorough evaluations and system implementation are remaining tasks for more advanced interactions.

7 CONCLUSION

We have presented our work on SoundScroll, a wristband input solution offloaded from the headset and leveraging finger friction sound to detect the sliding finger's contact state and speed. In our user study, we demonstrated that our system can provide fast detection of contact state every 20 ms with a high accuracy of 93.3%. Along with contact detection. SoundScroll also provides rich information on finger motion with speed regression using a single model. We showed the sound-based approach is a suitable wrist-worn solution compared to the IMU-based approach.

REFERENCES

[1] William Buxton et al. 1990. A three-state model of graphical input. In *Human-computer interaction-INTERACT*, Vol. 90. Citeseer, 449–456.

[2] Ke-Yu Chen, Shwetak N. Patel, and Sean Keller. 2016. Finexus: Tracking Precise Motions of Multiple Fingertips Using Magnetic Sensing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1504–1514. https://doi.org/10.1145/2858036.2858125

[3] Yi Fei Cheng, Tiffany Luong, Andreas Rene Fender, Paul Streli, and Christian Holz. 2022. ComforTable User Interfaces: Surfaces Reduce Input Error, Time, and Exertion for Tabletop and Mid-air User Interfaces. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 150–159. https://doi.org/10.1109/ISMAR55827.2022.00029

[4] Benoit Delhaye, Vincent Hayward, Philippe Lefèvre, and Jean-Louis Thonnard. 2012. Texture-induced vibrations in the forearm during tactile exploration. *Frontiers in behavioral neuroscience* 6 (2012), 37. https://www.frontiersin.org/articles/10.3389/fnbeh.2012.00037/full

[5] Nathan Devrio and Chris Harrison. 2022. DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 56, 13 pages. https://doi.org/10.1145/3526113.3545634

[6] S.R. Ellis, F. Breant, B. Manges, R. Jacoby, and B.D. Adelstein. 1997. Factors influencing operator interaction with virtual objects viewed via head-mounted see-through displays: viewing conditions and rendering latency. In *Proceedings of IEEE 1997 Annual International Symposium on Virtual Reality*. 138–145. https://doi.org/10.1109/VRAIS.1997.583063

[7] Stephen R Ellis, Mark J Young, Bernard D Adelstein, and Sheryl M Ehrlich. 1999. Discrimination of changes of latency during voluntary hand movement of virtual objects. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 43. SAGE Publications Sage CA: Los Angeles, CA, 1182–1186.

[8] Jun Gong, Aakar Gupta, and Hrvoje Benko. 2020. Acustico: Surface Tap Detection and Localization using Wrist-based Acoustic TDOA Sensing. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 406–419. https://doi.org/10.1145/3379337.3415901

[9] Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, and Charles C Kemp. 2024. PressureVision++: Estimating Fingertip Pressure from Diverse RGB Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 8698–8708.

[10] Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. 2022. Pressurevision: Estimating hand pressure from a single rgb image. In *European Conference on Computer Vision*. Springer, 328–345.

[11] Yizheng Gu, Chun Yu, Zhipeng Li, Weiqi Li, Shuchang Xu, Xiaoying Wei, and Yuanchun Shi. 2019. Accurate and Low-Latency Sensing of Touch Contact on Any Surface with Finger-Worn IMU Sensor. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1059–1070. https://doi.org/10.1145/3332165.3347947

[12] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, Asaf Nitzan, Gang Dong, Yuting Ye, Lingling Tao, Chengde Wan, and Robert Wang. 2020. MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality. *ACM Trans. Graph.* 39, 4, Article 87 (aug 2020), 13 pages. https://doi.org/10.1145/3386569.3392452

[13] Chris Harrison, Desney Tan, and Dan Morris. 2010. Skinput: Appropriating the Body as an Input Surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/1753326.1753394

[14] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 71 (jun 2020), 24 pages. https://doi.org/10.1145/3397306

[15] Ricardo Jota, Albert Ng, Paul Dietz, and Daniel Wigdor. 2013. How Fast is Fast Enough? A Study of the Effects of Latency in Direct-Touch Pointing Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2291–2300. https://doi.org/10.1145/2470654.2481317

[16] Wolf Kienzle and Ken Hinckley. 2014. LightRing: Always-Available 2D Input on Any Surface. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 157–160. https://doi.org/10.1145/2642918.2647376

[17] Wolf Kienzle, Eric Whitmire, Chris Rittaler, and Hrvoje Benko. 2021. ElectroRing: Subtle Pinch and Touch Detection with a Ring. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 3, 12 pages. https://doi.org/10.1145/3411764.3445094

[18] Daehwa Kim and Chris Harrison. 2022. EtherPose: Continuous Hand Pose Tracking with Wrist-Worn Antenna Impedance Characteristic Sensing. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 58, 12 pages. https://doi.org/10.1145/3526113.3545665

[19] David Kim, Otmar Hilliges, Shahram Izadi, Alex D. Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/2380116.2380139

[20] Scikit Learn. 2023. sklearn.ensemble.ExtraTreesClassifier. (2023). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[21] Chen Liang, Xutong Wang, Zisu Li, Chi Hsia, Mingming Fan, Chun Yu, and Yuanchun Shi. 2023. ShadowTouch: Enabling Free-Form Touch-Based Hand-to-Surface Interaction with Wrist-Mounted Illuminant by Shadow Projection. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 14 pages. https://doi.org/10.1145/3586183.3606785

[22] Chen Liang, Chun Yu, Yue Qin, Yuntao Wang, and Yuanchun Shi. 2021. DualRing: Enabling Subtle and Expressive Hand Interaction with Dual IMU Rings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 115 (sep 2021), 27 pages. https://doi.org/10.1145/3478114

[23] Robert W. Lindeman, John L. Sibert, and James K. Hahn. 1999. Towards Usable VR: An Empirical Study of User Interfaces for Immersive Virtual Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 64–71. https://doi.org/10.1145/302979.302995

[24] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D Hand Pose Tracking using EMG Wearables. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1471–1482. https://doi.org/10.1145/3442381.3449890

[25] I. Scott MacKenzie and Colin Ware. 1993. Lag as a Determinant of Human Performance in Interactive Systems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 488–493. https://doi.org/10.1145/169059.169431

[26] Manuel Meier, Paul Streli, Andreas Fender, and Christian Holz. 2021. TapID: Rapid Touch Interaction in Virtual Reality using Wearable Sensing. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. 519–528. https://doi.org/10.1109/VR50410.2021.00076

[27] Morph. 2023. Sensel Touchpad. (2023). https://sensel.com/

[28] Morph. 2024. Editing an Overlay. (2024). https://guide.sensel.com/app/#editing-an-overlay

[29] Adiyan Mujibiya, Xiang Cao, Desney S. Tan, Dan Morris, Shwetak N. Patel, and Jun Rekimoto. 2013. The sound of touch: on-body touch and gesture sensing based on transdermal ultrasound propagation. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces* (St. Andrews, Scotland, United Kingdom) *(ITS '13)*. Association for Computing Machinery, New York, NY, USA, 189–198. https://doi.org/10.1145/2512349.2512821

[30] Albert Ng, Julian Lepinski, Daniel Wigdor, Steven Sanders, and Paul Dietz. 2012. Designing for Low-Latency Direct-Touch Input. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 453–464. https://doi.org/10.1145/2380116.2380174

[31] Anh Nguyen and Amy Banic. 2014. 3DTouch: A wearable 3D input device with an optical sensor and a 9-DOF inertial measurement unit. *CoRR* abs/1406.5581 (2014). arXiv:1406.5581 http://arxiv.org/abs/1406.5581

[32] Takehiro Niikura, Yoshihiro Watanabe, and Masatoshi Ishikawa. 2014. Anywhere Surface Touch: Utilizing Any Surface as an Input Area. In *Proceedings of the 5th Augmented Human International Conference* (Kobe, Japan) *(AH '14)*. Association for Computing Machinery, New York, NY, USA, Article 39, 8 pages. https://doi.org/10.1145/2582051.2582090

[33] James F. O'Brien, Chen Shen, and Christine M. Gatchalian. 2002. Synthesizing Sounds from Rigid-Body Simulations. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (San Antonio, Texas) *(SCA '02)*. Association for Computing Machinery, New York, NY, USA, 175–181. https://doi.org/10.1145/545261.545290

[34] Ju Young Oh, Jun Lee, Joong Ho Lee, and Ji Hyung Park. 2017. Anywhere-touch: Finger tracking method on arbitrary surface using nailed-mounted imu

for mobile hmd. In *HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I 19*. Springer, 185–191.

[35] Farshid Salemi Parizi, Eric Whitmire, and Shwetak Patel. 2020. AuraRing: Precise Electromagnetic Finger Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 150 (sep 2020), 28 pages. https://doi.org/10.1145/3369831

[36] Andriy Pavlovych and Wolfgang Stuerzlinger. 2011. Target following performance in the presence of latency, jitter, and signal dropouts. In *Proceedings of Graphics Interface 2011*. 33–40.

[37] Gustavo Thebit Pfeiffer, Ricardo Guerra Marroquim, and Antonio Alberto Fernandes de Oliveira. 2014. WebcamPaperPen: A Low-Cost Graphics Tablet. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*. 87–94. https://doi.org/10.1109/SIBGRAPI.2014.54

[38] Helena Roeber, John Bacus, and Carlo Tomasi. 2003. Typing in thin air: the canesta projection keyboard - a new method of interaction with electronic devices. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI EA '03)*. Association for Computing Machinery, New York, NY, USA, 712–713. https://doi.org/10.1145/765891.765944

[39] Xiyuan Shen, Chun Yu, Xutong Wang, Chen Liang, Haozhan Chen, and Yuanchun Shi. 2-24. MouseRing: Always-available Touchpad Interaction with IMU Rings. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

[40] Yilei Shi, Haimo Zhang, Kaixing Zhao, Jiashuo Cao, Mengmeng Sun, and Suranga Nanayakkara. 2020. Ready, Steady, Touch! Sensing Physical Contact with a Finger-Mounted IMU. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 59 (jun 2020), 25 pages. https://doi.org/10.1145/3397309

[41] Adalberto L. Simeone, Eduardo Velloso, and Hans Gellersen. 2015. Substitutional Reality: Using the Physical Environment to Design Virtual Reality Experiences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3307–3316. https://doi.org/10.1145/2702123.2702389

[42] Anton R Sobinov and Sliman J Bensmaia. 2021. The neural mechanisms of manual dexterity. *Nature Reviews Neuroscience* 22, 12 (2021), 741–757.

[43] Sonion. 2023. VOICE PICK UP BONE SENSOR (VPU). (2023). https://www.sonion.com/hearing/bone-conduction-sensors-and-actuators/vpu-voice-pick-up-sensor/.

[44] Paul Streli, Jiaxi Jiang, Andreas Rene Fender, Manuel Meier, Hugo Romat, and Christian Holz. 2022. TapType: Ten-Finger Text Entry on Everyday Surfaces via Bayesian Inference. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 497, 16 pages. https://doi.org/10.1145/3491102.3501878

[45] Paul Streli, Jiaxi Jiang, Juliete Rossie, and Christian Holz. 2023. Structured Light Speckle: Joint Ego-Centric Depth Estimation and Low-Latency Contact Detection via Remote Vibrometry. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/3586183.3606749

[46] Paul Streli, Jiaxi Jiang, Juliete Rossie, and Christian Holz. 2023. Structured Light Speckle: Joint Ego-Centric Depth Estimation and Low-Latency Contact Detection via Remote Vibrometry. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/3586183.3606749

[47] Sarah Elizabeth Tomlinson. 2009. *Understanding the friction between human fingers and contacting surfaces.* Ph. D. Dissertation. University of Sheffield.

[48] Kees van den Doel, Paul G. Kry, and Dinesh K. Pai. 2001. FoleyAutomatic: Physically-Based Sound Effects for Interactive Simulation and Animation. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. Association for Computing Machinery, New York, NY, USA, 537–544. https://doi.org/10.1145/383259.383322

[49] Radu-Daniel Vatavu and Laura-Bianca Bilius. 2021. GestuRING: A Web-Based Tool for Designing Gesture Input with Rings, Ring-Like, and Ring-Ready Devices. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 710–723. https://doi.org/10.1145/3472749.3474780

[50] Nicolas Villar, Shahram Izadi, Dan Rosenfeld, Hrvoje Benko, John Helmes, Jonathan Westhues, Steve Hodges, Eyal Ofek, Alex Butler, Xiang Cao, and Billy Chen. 2009. Mouse 2.0: multi-touch meets the mouse. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology* (Victoria, BC, Canada) *(UIST '09)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/1622176.1622184

[51] Anandghan Waghmare, Youssef Ben Taleb, Ishan Chatterjee, Arjun Narendra, and Shwetak Patel. 2023. Z-Ring: Single-Point Bio-Impedance Sensing for Gesture, Touch, Object and User Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 150, 18 pages. https://doi.org/10.1145/3544548.3581422

[52] Dan Wu, Ruiyang Gao, Youwei Zeng, Jinyi Liu, Leye Wang, Tao Gu, and Daqing Zhang. 2020. FingerDraw: Sub-Wavelength Level Finger Motion Tracking with WiFi Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 31 (mar 2020), 27 pages. https://doi.org/10.1145/3380981

[53] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M. Kitani. 2020. Back-Hand-Pose: 3D Hand Pose Estimation for a Wrist-Worn Camera via Dorsum Deformation Network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1147–1160. https://doi.org/10.1145/3379337.3415897

[54] Robert Xiao, Julia Schwarz, Nick Throm, Andrew D Wilson, and Hrvoje Benko. 2018. MRTouch: Adding touch input to head-mounted mixed reality. *IEEE transactions on visualization and computer graphics* 24, 4 (2018), 1653–1660.

[55] Chenhan Xu, Bing Zhou, Gurunandan Krishnan, and Shree Nayar. 2023. AO-Finger: Hands-Free Fine-Grained Finger Gesture Recognition via Acoustic-Optic Sensor Fusing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 306, 14 pages. https://doi.org/10.1145/3544548.3581264

[56] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. 2012. Magic Finger: Always-Available Input through Finger Instrumentation. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 147–156. https://doi.org/10.1145/2380116.2380137

[57] H. Zahouani, R. Vargiolu, G. Boyer, C. Pailler-Mattei, L. Laquièze, and A. Mavon. 2009. Friction noise of human skin in vivo. *Wear* 267, 5 (2009), 1274–1280. https://doi.org/10.1016/j.wear.2009.03.007 17th International Conference on Wear of Materials.

[58] Yang Zhang, Wolf Kienzle, Yanjun Ma, Shiu S. Ng, Hrvoje Benko, and Chris Harrison. 2019. ActiTouch: Robust Touch Detection for On-Skin AR/VR Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1151–1159. https://doi.org/10.1145/3332165.3347869

[59] Yuliang Zhao, Xianshou Ren, Chao Lian, Kunyu Han, Liming Xin, and Wen J Li. 2021. Mouse on a Ring: A Mouse Action Scheme Based on IMU and Multi-Level Decision Algorithm. *IEEE Sensors Journal* 21, 18 (2021), 20512–20520.