SAWSense: Using Surface Acoustic Waves for Surface-bound Event Recognition

Yasha Iravantchi
University of Michigan
Ann Arbor, Michigan, USA
yiravan@umich.edu

Yi Zhao
University of Michigan
Ann Arbor, Michigan, USA
eve.yizhao@gmail.com

Kenrick Kin
Meta Reality Labs
Redmond, Washington, USA
kenrickkin@meta.com

Alanson P. Sample
University of Michigan
Ann Arbor, Michigan, USA
apsample@umich.edu

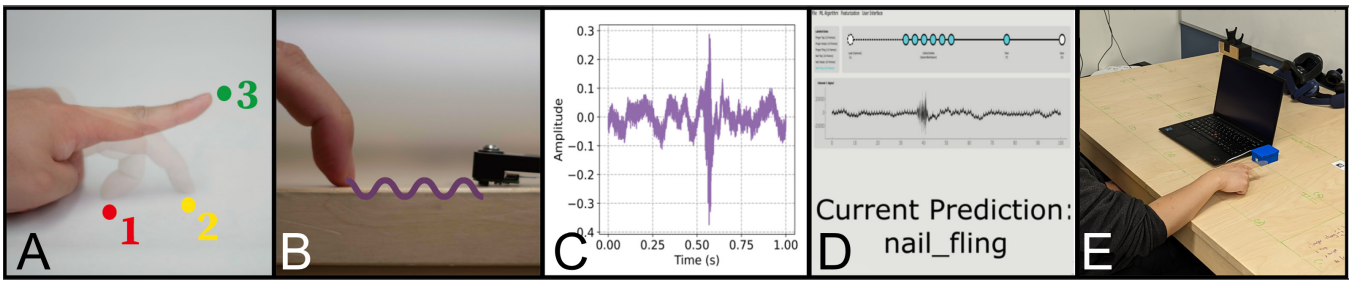

Figure 1: A conceptual depiction of the SAWSense sensing system. A user presses their nail against the table and performs a fling gesture (Panel A). When the finger contacts the table, a Surface Acoustic Wave is generated and propagates along the surface-to-air boundary and is sampled by a Voice PickUp Unit (VPU) (Panel B). The output of the VPU is converted and depicted in Panel C. The gesture is then classified with our machine learning pipeline (Panel D) and, as one possible application, is integrated into the feet of a laptop to extend the interaction area to the surface (Panel E).

ABSTRACT

Enabling computing systems to understand user interactions with everyday surfaces and objects can drive a wide range of applications. However, existing vibration-based sensors (e.g., accelerometers) lack the sensitivity to detect light touch gestures or the bandwidth to recognize activity containing high-frequency components. Conversely, microphones are highly susceptible to environmental noise, degrading performance. Each time an object impacts a surface, Surface Acoustic Waves (SAWs) are generated that propagate along the air-to-surface boundary. This work repurposes a Voice PickUp Unit (VPU) to capture SAWs on surfaces (including smooth surfaces, odd geometries, and fabrics) over long distances and in noisy environments. Our custom-designed signal acquisition, processing, and machine learning pipeline demonstrates utility in both interactive and activity recognition applications, such as classifying trackpad-style gestures on a desk and recognizing 16 cooking-related activities, all with >97% accuracy. Ultimately, SAWs offer a unique signal that can enable robust recognition of user touch and on-surface events.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580991

CCS CONCEPTS

• **Human-centered computing** → **Interaction devices**; *Ubiquitous and mobile devices*; **Gestural input**; • **Hardware** → *Sensor devices and platforms.*

KEYWORDS

Surface Acoustic Wave, Touch Detection, Activity Recognition, Gesture Interface, Acoustics, Sensing

ACM Reference Format:
Yasha Iravantchi, Yi Zhao, Kenrick Kin, and Alanson P. Sample. 2023. SAWSense: Using Surface Acoustic Waves for Surface-bound Event Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3580991

1 INTRODUCTION

Effective means of enabling computing systems to sense and understand user gestures, activities, and the context in which objects are used can enable a wide range of interactive and assistive technologies. Common approaches for sensing users' actions have relied on cameras and computer vision techniques to capture gestures and estimate pose, as well as microphones for speech recognition and

audio event classification. While these methods are effective in well-defined applications, they lack the fidelity and sensitivity to sense many surface-based events. For example, depth cameras struggle to resolve hover vs. touch events, and microphones cannot "hear" very quiet surface events. Additionally, cameras and microphones are often susceptible to issues in real-world environments, such as line-of-sight interference and background noise, such as speech or music, which degrade the performance of their applications.

To address these shortcomings, researchers have explored various methods for integrating sensing mechanisms into planar and curved surfaces to better capture the precise timing and nature of user and object interaction events. Examples include capacitive sensing arrays [13, 23, 73], conductive paint [72], and resistive sensor arrays [1, 15, 29], which are integrated into the surface of the object. Additionally, optical methods have been explored using arrays of break beam sensors [27, 47], and multiple time-of-flight sensors [9, 68]. To overcome the high instrumentation costs of covering surfaces with dense sensors, researchers have investigated methods for detecting the physical vibrations caused when a user's fingers or an object touches a given surface, thus requiring only one or two instrumentation points. However, acoustic sensing methods using microphones that rely on measuring propagating sound waves are easily overwhelmed by ambient environmental noise and thus must typically limit their sensing bandwidth to a few kilohertz [24]. Likewise, geophones and accelerometers that measure mechanical vibrations also have low sampling bandwidth (less than 250Hz), and piezoelectric-based contact microphones are designed to resonate only at the design frequency of 2-4kHz and thus are best used to detect impulses over relatively short distances. This bandwidth limitation ultimately restricts the ability of these approaches to support various applications across different environments.

To overcome issues related to sound and vibration-based sensing on surfaces, this paper investigates Surface Acoustic Waves (SAWs), methods to capture these signals for robust gesture interfaces, and the detection of human-object interaction events with minimal instrumentation overhead. Similar to sound and mechanical vibration, Surface Acoustic Waves are generated when an object, such as a user's finger, makes contact with a surface. However, unlike sounds (which propagate through the air) and mechanical vibrations (which propagate through the bulk of the medium), SAWs are bound to the surface-to-air boundary and propagate as 2D waves along the surface of an object. Furthermore, SAWs have several unique propagation characteristics, such as low attenuation and immunity to sound waves, permitting them to be captured at longer distances and in noisy environments, ideal for surface-based sensing.

To take advantage of SAWs, we repurpose a Voice PickUp Bone Sensor (VPU) [11] which was originally developed for earbuds to capture waves traveling from the vocal cords to the tissues in the inner ear, allowing a person wearing them to speak and be heard clearly in a crowded and noisy environment. These sensors are hermetically sealed, allowing them to reject sound waves through the air and capture SAWs only through contact. Importantly, they are fabricated using a MEMS process, granting them the bandwidth and physical footprint typical to traditional MEMS microphones found in commercial off-the-shelf devices. Through an initial set of experiments, we found these VPUs capture 96.9% more information power than accelerometers/geophones and maintain a similar bandwidth to traditional microphones while being robust to environmental sounds. These VPUs form the inputs to the SAWSense pipeline, which features signal processing and machine learning optimized for Surface Acoustic Waves.

To illustrate the breadth of SAWs and their utility for various interactive applications, SAWSense was evaluated across two domains of HCI research: gestural input and ubiquitous activity recognition. On traditional flat surfaces, such as on a desk, SAWSense recognizes trackpad-style gestures and, with the aid of a second VPU sensor, can also infer the direction of these gestures. In the home environment, SAWSense can perform activity recognition tasks and robustly classify 16 different cooking-related events, even in the presence of noise. In all of these evaluations, SAWSense achieved >97% classification accuracy. To illustrate additional future avenues of research, SAWSense explored SAWs on odd geometries, such as playful interactions with a dragon toy and 3D-printed tangibles – even on a scaled-back pipeline that can run on microcontrollers. Additionally, we used SAWSense to explore interactions on fabrics, such as a jacket sleeve, showing SAWs are not limited to rigid surfaces. While there is no one-size-fits-all sensing approach, SAWs demonstrate a unique set of properties that make them practical for various surface-bound tasks. Overall, while prior work has explored vibroacoustic-based contact sensing, we show that SAWs are a practical sensing approach that is low cost, low power, has a small profile, is robust to ambient noise, and can work across surfaces of different materials and shapes, making SAWSense a more practical approach for surface-acoustic sensing.

This paper makes the following contributions to the use of Surface Acoustic Wave sensing for surface event recognition:

(1) A new sensing modality for capturing Surface Acoustic Waves (SAWs)
(2) An information power analysis used to optimize SAW signal processing and machine learning performance
(3) A system that can robustly detect SAW-based gestures and events in acoustically challenging situations
(4) A data augmentation and feature transformation pipeline that enhances cross-users and cross-material accuracy
(5) A demonstration of the breath of SAW sensing applications on non-traditional surfaces

The remainder of the paper is organized in the following manner: Section 2 presents a review of relevant literature to contextualize SAWSense with other touch and object recognition sensing methods. Section 3 provides a brief introduction to Surface Acoustic Waves, followed by a series of technical comparisons between standard sensors and VPUs. Section 4 describes the hardware implementation of SAWSense and the design and validation of the signal processing and machine learning pipeline. Section 5 evaluates SAWSense in a range of usage domains, demonstrating the breadth of Surface Acoustic Waves that are useful for classifying surface-related events. Section 6 details potential avenues for future works and discusses the results and limitations of SAWSense.

2 RELATED WORKS

This section contextualizes SAWSense's contribution along three axes: the sensor type (passive contact-based sensing approaches),

the signal processing and machine learning approach (acoustic event detection), and the application domain (surface interaction). We now describe each area relative to SAWSense in further detail.

2.1 Passive Sensing for Contact-based Interaction

Making surfaces interactive to touch and contact is a well-established area of research in the HCI literature. One approach is to integrate sensing into the surface of interest, such as a wall or object, through conductive paint [72], capacitive sensing arrays [13, 23, 73], and resistive sensing arrays [1, 15, 29]. For example, Touch and Activate [49] use piezotransducers to inject acoustic signals into an object to determine touch events. However, these systems are all active sensing systems and require either the entire surface to be treated or multiple instrumentation points. Passive sensing approaches do not utilize an active signal, which can potentially improve the ease of instrumentation. For example, IDSense [40] requires only the addition of a single RFID tag on an object to enable touch detection. However, that object must be within range of an RFID reader. The predominant optical approach is depth-image-based methods [7, 65–67], which drive interactive experiences such as RoomAlive [33] and WorldKit [69]. Other optical approaches where a camera-style device is placed in the environment include LaserWall [51], which uses a laser rangefinder, and HeatWave, [38] which places a thermal camera above a surface and uses heat transfer from the hand to determine touch events. Finally, other optical approaches to determine touch events include arrays of break beam sensors [27, 47] and time-of-flight sensor arrays [9, 68].

Most similar to SAWSense are vibroacoustic touch sensing approaches [22, 60, 71] that use IMUs [21, 32], geophones [28, 50], piezodisks/contact microphones [8, 41], and traditional microphones [24, 25, 31, 37]. While accelerometers and geophones are robust to environmental sound, they have bandwidth limitations (typically less than 250Hz) and cannot sense higher-frequency events. Piezodisks, which are resistant to external sounds, typically have a design (resonant) frequency and provide very small signals outside of those frequencies. For greater overall sensing bandwidth, systems can utilize traditional microphones but must incorporate high-pass filters since background noise from the environment can pollute the incoming signal and degrade performance [24, 26]. Thus there is a need for a wide-band sensing approach that is robust to external sounds for contact-based sensing.

Recently, novel and specially designed Voice Pick Up (VPU) sensors have been developed to capture a speaker's voice through contact with the skin without capturing environmental sounds. SAWSense repurposes these sensors to capture Surface Acoustic Waves, which cannot be induced by external sounds. These devices maintain the rich signal qualities of geophones (with greater bandwidth to effectively represent speech) while maintaining a small MEMS footprint similar to IMUs, enabling easy integration into devices. These devices have only very recently entered the academic literature for bone-conduction-enhanced speech recognition [39] and advanced hearing aid research [54]. Since the VPUs sensors offer new capabilities, we provide a comparison to accelerometers, geophones, and traditional microphones in Section 3.

2.2 Acoustic Event Detection

While IMUs and geophone-based sensing offer an established approach for detecting surface interaction events with robustness to environmental noise, they have relatively narrow bandwidth [12, 61] compared to traditional microphones [2] making them unable to fully take advantage of existing multi-class acoustic sound recognition (ASR) and sound event detection (SED) pipelines. Alternatively, while traditional microphones have a well-established ability for use in multi-class sound classification, the sounds of surface-based gestures overlap in frequency with speech, music, and other environmental sounds, presenting a challenge for robust operation in real-life environments. The VPU offers similar bandwidth to traditional microphones, providing an opportunity to leverage conventional ASR/SED techniques, all while having the environmental robustness of IMUs and geophones.

Traditional ASR and SED pipelines typically convert time-domain acoustic waves into features like Fast Fourier Transforms (FFTs) and Mel-frequency Cepstral Coefficients (MFCCs). These features enable machine learning algorithms to enhance human speech features, reduce computation complexity, improve generalizability, and homogenize the inputs to an ML pipeline [57]. For example, Piczak et al. [53] and NELS [17] both use 60 mel-bands to extract features for environmental sound classification. Bello et al. provide a deeper explanation of MFCCs and a survey on urban sound featurization, augmentation, and classification techniques for traditional microphone audio [6]. However, MFCCs may not be the optimal feature for non-conventional audio. For instance, PrivacyMic uses log-bin FFTs to distribute more features at lower ultrasonic frequencies [30]. Since featurizing passive wide-band SAW signals is underexplored, SAWSense uses an information-power approach to design a custom MFCC where the location and range of the filter banks are optimized for SAWs instead of human sounds.

Beyond traditionally crafted features, systems such as SoundNet [4] and AudioSet [20] have leveraged large datasets to train Convolutional Neural Nets (CNNs) that can process raw audio waveforms directly, bypassing the need for hand-crafted features. In the absence of large datasets, these approaches benefit from sound augmentation, which rapidly generates synthetic data with added noise and effects (e.g., pitch shift, reverb, time stretch) from relatively small datasets, to improve the recognition performance and robustness of their models [6, 58]. While SAWSense ultimately uses traditional feature modalities and much smaller machine learning models, our preliminary results show that augmentation methods are compatible with SAWs, improving across-user performance for trackpad-style gestures. Furthermore, Section 5 shows that training data collected on one material type can be transferred to a different material type using a non-linear transform.

2.3 Surface Interaction

In ubiquitous and aware-home applications, previous works have explored hand-to-surface applications, including Direct [70] and WorldKit [69]. With emerging AR/VR input needs, works have explored placing virtual keyboards on physical surfaces as haptic sources and demonstrate the capacity for on-table virtual keyboard typing via camera systems [43, 56]. These camera-based works acknowledge that there are drawbacks to optical approaches to tracking the placement of fingers relative to surfaces. As a workaround,

these systems place a virtual plane with a fixed distance above the physical surface where the finger has to cross to register as a tap, effectively requiring the hands to hover above the surface unlike traditional physical inputs. However, it is especially important for immersion into AR/VR experiences to maintain traditional physical input modalities and incorporate tactile feedback so that the user can employ affordances from the real world in the virtual world.

Recent works like Acoustico [22] and TapID [45] specifically explore using wearables to aid in surface tap detection for AR/VR environments. These devices work in concert with optical-based hand tracking, which provides a general location of the hands, and these systems precisely detect touch events. Similarly, SAWSense explores improving surface tap and gesture detection, but instruments the surface of interest directly with a single sensor. Additionally, SAWSense explores these touch events on objects and clothes, as detailed in Section 5 and Section 6.

3 SENSOR COMPARISONS

This section provides a brief introduction to Surface Acoustic Waves (SAWs) and their wave characteristics, followed by a series of comparisons between three commonly used sensors (accelerometers, microphones, and geophones) and a Voice PickUp (VPU) sensor. The first comparison is a frequency response evaluation, measuring the range of frequencies that these sensors can reliably capture. A distance response evaluation compares the sensing range of each sensor. A qualitative evaluation details the behavior of these sensors in a variety of conditions, such as in a noisy environment and on different surface materials. Finally, a small-scale desk-event classification task compares the relative information power and machine learning (ML) performance for common office events when using each of the four sensors. Through each evaluation, VPUs show qualities that are ideal for a wide variety of surface-bound sensing tasks, matching or outperforming the other three sensors.

3.1 Brief Introduction to SAWs

When an object impacts a surface, such as a table, the object initiates a transfer of kinetic energy into the material, which launches 1) an acoustic wave that travels through the air as sound, 2) a 3-dimensional mechanical wave that travels through the bulk of the medium (e.g., a table) as vibrations, and 3) a 2-dimensional Surface Acoustic Wave which is coupled to the surface-to-air boundary of the object and thus propagates along the surface. SAWs are a sub-class of acoustic waves that are an amalgamation of several propagation mechanisms: Rayleigh waves that include both longitudinal and transverse propagation components and Love waves that are horizontally polarized surface waves [10].

The propagation characteristics of Surface Acoustic Waves offer unique advantages as a sensing modality. Most notably, since SAWs are coupled to a two-dimensional surface, their in-plane amplitude decays at a rate of $1/\sqrt{r}$ (where r is the radial distance from the source of impact). Compared to bulk mechanical waves (i.e., vibrations) that propagate in three dimensions as they travel through the body of the object and decay at a rate of $1/\sqrt[3]{r}$, SAWs propagate for longer distances. As shown in Section 3.4, SAWs can be detected nearly anywhere on the surface of a table due to their low rate of attenuation. Additionally, SAWs are only generated through direct physical coupling, and there is no transfer mechanism from free-air acoustic signals to surface acoustic signals. As a result, sounds in the environment (e.g., speech, music) do not interfere with SAWs propagating along an object's surface. This is particularly advantageous for sensing in real-world noisy environments, which is a challenge for traditional microphones since background noise often overlaps with the signals of interest in audio-based sensing tasks.

Historically, Surface Acoustic Waves have been studied in the field of seismology, where earthquakes generate low-frequency SAWs that propagate along the surface of the earth and can be used to help infer underground geological structures [3]. Works have leveraged the active use of SAWs for flow measurement [35], precision droplet direction in microfluidics [19], and tactile displays [48]. Active SAW sensors are becoming more commercially available, but operate as tuned narrowband emitter/receiver pairs with MHz to GHz center frequencies [16]. Within the HCI literature, Swaminathan et al. [63] use a 174MHz emitter/receiver pair to instrument surfaces. An in-depth explanation and analysis of SAWs can be found in [10]. For reference, SAWSense is not an active SAW sensing approach and is passively sensitive to a wide range of frequencies in the kHz range (roughly 0-20kHz) as opposed to narrowband frequencies in the MHz and GHz ranges.

More recently, Voice PickUp (VPU) sensors have been developed to capture the waves that propagate from a person's vocal cords to the tissues in the ear canal while speaking, allowing for in-ear headphones to capture speech reliably even in the presence of significant environmental noise. Importantly, to faithfully capture speech sounds, these sensors need to have a relatively wide bandwidth. Additionally, these devices are hermetically sealed and do not capture acoustic waves from the air (i.e., sound); there is no acoustic leakage when the vent hole is sealed after the reflow process [11]. As a result, the qualities that enable VPUs to effectively capture speech in-ear without capturing external sounds are also useful for a wide range of sensing tasks evaluated in this work.

3.2 Sensor Selection

As detailed in the Related Works section, there are three commonly used sensors for surface-based activity and event detection: the accelerometer, microphone, and geophone. While there are many accelerometers to choose from, the STMicroelectronics LSM9DS1 [42] is a readily available accelerometer and is found in the Arduino Nano 33 Sense. It can sample reliably up to 500Hz and is generally representative of the kinds of higher-end accelerometers found in commodity devices. For our microphone, we selected the MiniDSP UMA-8 [64], which uses 7 low-noise, low-distortion, omnidirectional microphones [62] and is representative of the types of microphones found in many smartphones and smart speakers. We set the UMA-8 to "raw output" mode to avoid introducing additional factors in this evaluation and selected the channel facing in the "forward" direction, which is sampled at 48kHz. For the geophone, we selected the ION Inc. SM-24 [61], which is the same model as the ones used in many prior HCI works, such as SurfaceVibe [50]. We paired the geophone with a Behringer UM2 [5] as our Analog-to-Digital Converter (ADC), which we confirmed to have a flat frequency response from 0.01Hz to 24kHz using a function generator [36], and set the sampling rate to 48kHz. Finally,

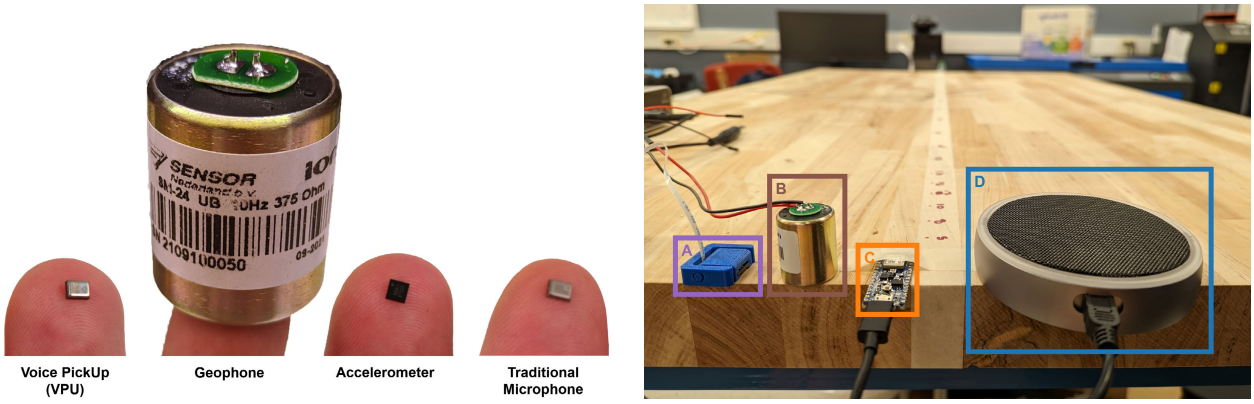

Figure 2: On the left, the relative size of each standalone sensor compared to a finger. On the right, the placement of the VPU (A), geophone (B), accelerometer (C), and traditional microphone (D) on a wooden workbench for the frequency response, distance response, and classification task evaluations. We note that while (D) contains 7 microphones, which increases the final size of the device but has no adverse effects on sensing ability, only one microphone is used in the evaluation.

we set our VPU-based sensor to a sampling rate of 48kHz. All devices were connected to a laptop via USB, and a script was run to simultaneously capture synchronized samples. All evaluations in this section were conducted in a large, quiet room (roughly 7.5m X 6m) with the workbench in the center of the room and no other occupants. The room did not have noticeable echos and contained office chairs, other workbenches, and fabrication tools (e.g., laser cutter, 3D Printer) that were not in operation. The sensors and their placement on the workbench can be seen in Figure 2.

3.3 Frequency Response Evaluation

Frequency response offers a metric to better understand the breadth of frequencies that these four sensors can reliably capture. In order to compute the frequency response for each device, a speaker was used to generate the three types of waves (vibrations, sounds, and SAWs) across a wide range of frequencies for each sensor. The speaker's driver generates sound for the microphone, and the movement of the driver causes the housing of the speaker to impact the surface at a given frequency, creating both free vibrations and SAWs. We selected a reasonably sized single speaker (GigaWorks T20) and set the speaker driver-down on a wooden workbench. A function generator [36] provided input to the speaker, performing a linear sweep from 1Hz to 20kHz. We visually verified the speaker's driver oscillating at the lower frequencies and confirmed the speaker's sound output at 20kHz using a Dodotronic 384kHz microphone [14], confirming no internal filters limited the speaker's output. The volume of the speaker was set to be sufficient for the housing to impact the surface and generate vibrations and SAWs but without causing the microphone to experience clipping/saturation. The sensors were placed 36cm away (the minimum distance that did not cause saturation/clipping across all four sensors) from the speaker and captured the sweep simultaneously. We note that this experiment was conducted in quiet conditions so as to not unreasonably hamper the regular microphone's performance.

We find a generally flat frequency response below 100Hz for the accelerometer, but its overall bandwidth (250Hz) is limited by the sampling rate of the sensor. For the microphone, we find a steep drop after 10kHz, which largely matches the manufacturer's datasheet. This attenuation for higher frequencies, however, does not mean the microphone cannot capture higher-frequency signals at all. The UMA-8, under these experimental conditions, captured the signals distinctly above the noise floor. This attenuation may not ultimately impede the microphone's ability to sense hand activities since Braun et al. [8] only observed signals up to 12kHz using a microphone. While the geophone is sampled at 48kHz, we see a steep drop in the response above 400Hz. We note, however, that the geophone's datasheet claims 240Hz bandwidth, frequencies above which are stated to be spurious. Faber et. al provides a deeper explanation of spurious frequencies [18]. Finally, the VPU maintains a relatively flat response, with a peak just before 10kHz. Figure 3 (left) shows the normalized frequency response curves for all four devices. Overall, from a bandwidth perspective, the VPU maintains significantly more bandwidth than accelerometers/geophones (almost 40x greater), allowing it to capture a richer set of frequencies more akin to that of regular microphones. Ultimately, we expect the VPU's significant bandwidth to better resolve a greater variety of events than accelerometers and geophones.

3.4 Distance Response Evaluation

While bandwidth considerations are important, the distance response defines the overall sensing range of these devices. To evaluate signal fidelity across distances, a small glass marble (5.5 grams, Dia=1.5cm) was dropped from a height of 5cm on a wooden workbench to create a standardized and repeatable impulse. We marked out distances logarithmically at 5, 8, 10, 13, 23, 28, 46, 56, 69, 84, 104, and 130 cm and recorded the peak amplitude of the impulse at each distance for each sensor. Figure 3 (right) shows the relative signal peak in dB with a -3dB reference line. We reiterate that this experiment was conducted in quiet conditions so as to not unreasonably hamper the regular microphone's performance.

While all sensors were able to capture the impulse with a signal strength above -12dB at 130cm, only the geophone and VPU were able to maintain above -3dB at 84cm, almost double the distance vs.

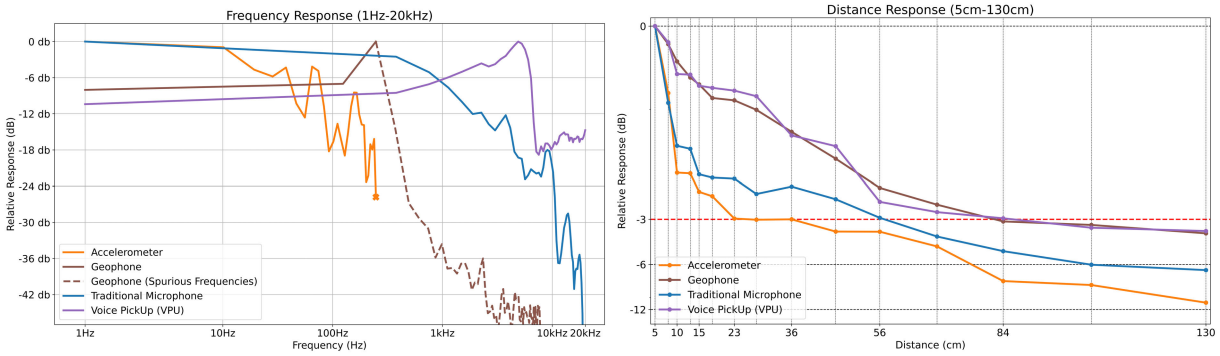

Figure 3: On the left, the frequency response curves from 1Hz to 20kHz are shown for the four sensors. The Nyquist limit of the accelerometer is denoted with an "X" and the spurious frequencies per the geophone's datasheet are denoted by a dashed line. On the right, the distance response curves from 5cm to 130cm are shown for the four sensors. Overall, we observe only the microphone and VPU have bandwidth in the kilohertz ranges and only the geophone and VPU maintain -3dB response at 84cm.

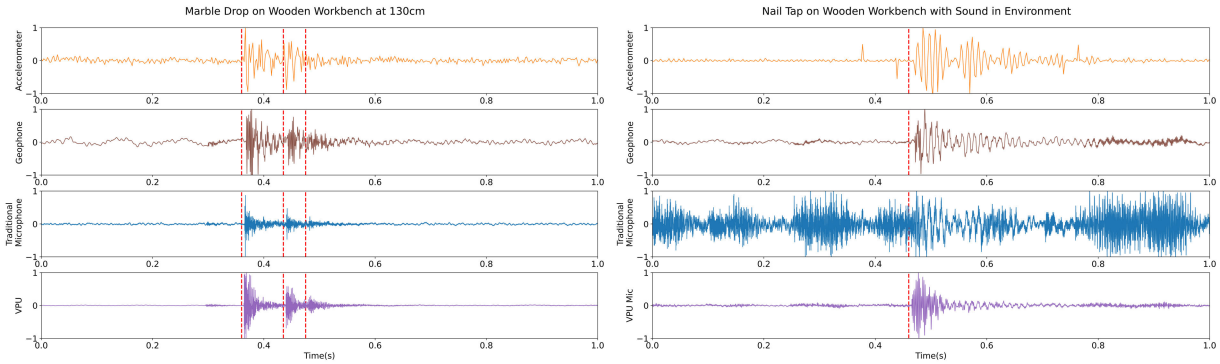

Figure 4: The plots on the left show the raw signal from the four sensors when a marble is dropped on a wooden workbench from a distance of 130cm. The dashed vertical red lines denote the three bounces of the marble. The plots on the right show the raw signals from the four sensors when a fingernail is used to tap the workbench, denoted by the dashed vertical line, while there is sound playing in the environment. Only the VPU has the sensitivity to capture the third bounce of the marble and be robust to external sounds.

the accelerometer. Additionally, peak signal strength is only one facet of overall signal fidelity; when we examine the raw signals at 130cm (see Figure 4, left), we observe that only the regular microphone and VPU are able to capture the third bounce at 130cm; the signal is too small to resolve above the noise floor for the accelerometer and the ringing inherent to the mass-spring system of the geophone dominates. These results indicate that it is possible to instrument a single point and capture the signal from interaction events over an entire table's surface using the VPU.

3.5 Environment and Situational Comparisons

The frequency and distance response comparisons show the signal fidelity of the four sensors in relatively ideal conditions. However, other factors reveal particular advantages and disadvantages for each sensor. This section details the behavior of the four sensors in common real-world environmental and situational conditions.

An important factor for a sensor is its robustness to various sources of environmental noise. We evaluated the ability of all four sensors to pick up events on a wooden workbench when everyday sounds (e.g., speech, music) were introduced to the environment as noise. For the microphone, the noise was captured along with the sound from the event, whereas the other three sensors did not capture the noise. Figure 4 (right) provides an example where a fingernail tapped a wooden workbench while NPR News Now played on a speaker in the room at a normal listening volume.

An additional factor is how well these sensors can operate on a wide variety of surfaces. Typically, most materials can reliably generate sounds, vibrations, and SAWs from impulses (e.g., placing an item on a surface), but vibrations/sounds/SAWs generated from friction (e.g., moving a mouse, swiping with a finger) are highly dependent on the surface. For example, on a wooden workbench, all but the accelerometer could resolve both a finger tap and a finger swipe on the surface. However, on significantly smoother surfaces, such as glass and granite, only the VPU could resolve the finger swipe above the noise floor. Figure 5 shows examples using four everyday materials: a textured wooden workbench, a smooth Ikea Linnmon tabletop, a granite countertop, and a glass coffee table.

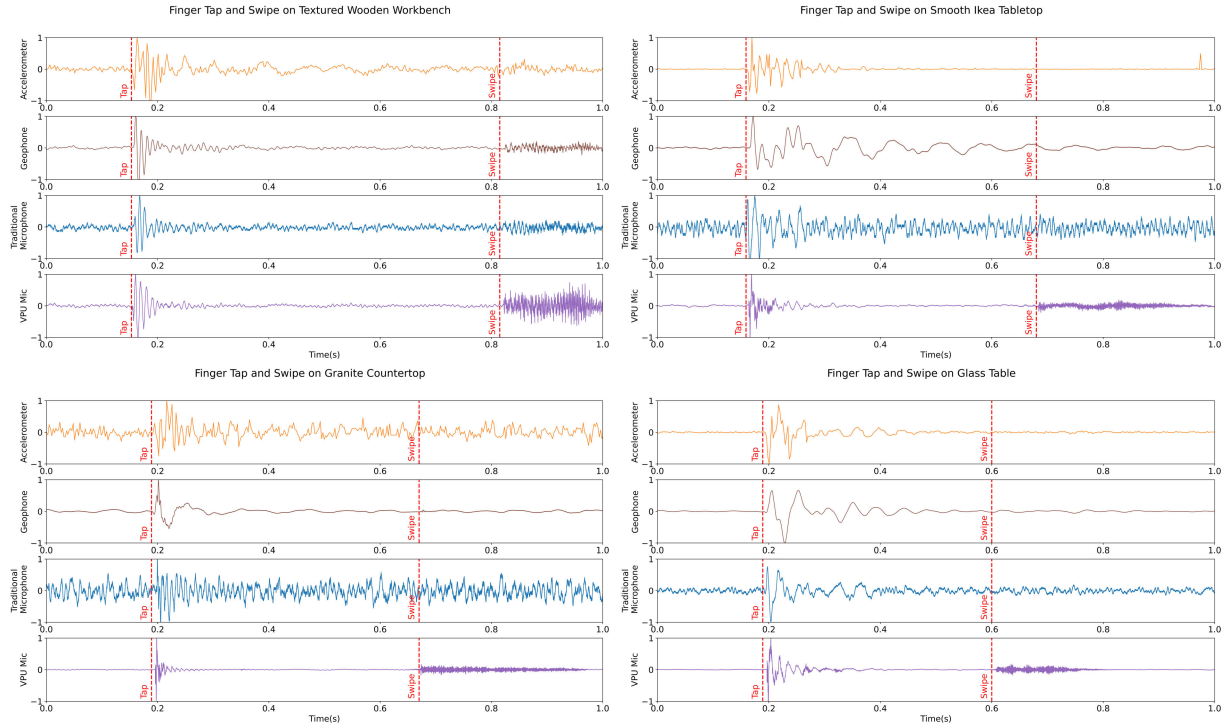

Figure 5: The plots show a finger tap followed by a finger swipe on four everyday surfaces. The beginning of each event is labeled and denoted by a dashed red vertical line. While all but accelerometers have visible signals above their noise floor from swipes on the textured wooden surface, only the VPU visibly captures swipes on the smoother surfaces.

Finally, the ability to integrate these sensors into devices and on various surfaces is a significant factor in whether the approaches will be adopted in real-world applications. Geophones, which have demonstrated significant sensing capabilities, are limited by their relatively large size (H=32mm, Dia=25mm, Weight=74g), cost ($65), and requirement for the mass-spring to be oriented in the direction of gravity (i.e., can only be placed upright). Conversely, cheap (often <$1 at scale) and small (see Figure 2 for size reference) Micro-ElectroMechanical Systems (MEMS) versions of accelerometers and microphones have found themselves integrated into numerous commodity devices. They are also orientation invariant, allowing them to be placed virtually anywhere on a surface or object. Since the VPU is manufactured using a MEMS process, it is similarly small, cost-effective, orientation invariant, and embodies almost all the integration advantages of other MEMS sensors.

3.6 Desk-related Event Evaluation

While the previous sections provide metrics that help inform the utility of a sensor for a set of tasks, we also wanted to evaluate whether these characteristics ultimately contribute to improved performance in a classification task. For this evaluation, we placed the four sensors on a wooden workbench in a relatively quiet environment to not unreasonably hamper the regular microphone. Then at a distance of 56cm (roughly half the length of a standard office desk), we performed a series of common desk-related events: placing a bottle, typing on a keyboard, placing a set of keys, moving a mouse, placing a smartphone, writing on a sheet of paper with a pencil, tapping a finger (without the fingernail), and swiping a finger (without the fingernail). We also collected samples of the "nothing" class, where no event took place. We collected 10 1-second instances of each, forming a round. We collected 10 rounds, resulting in 100 total instances for each class. In order to be as "fair" as possible when featurizing the signals from each sensor, we performed a Fast Fourier Transform (FFT) at 1Hz resolution (including the DC component), resulting in 251 features for the accelerometer (500Hz sampling rate) and 24001 features for the microphone, geophone, and VPU (48kHz sampling rate). Additionally, given the differences in sensor bandwidth and the number of features, we opt for an embedded classification method that integrates feature selection and maximizes the amount of information in the final feature set used in classification. Thus a Random Forest classifier (Scikit-Learn [59], default parameters) performs a 10-round cross-validation, where we train on nine rounds and test on a 10th for all combinations, reporting the average results. We found the accelerometer achieved 69.8% (SD = 2.5%), the traditional microphone at 81.8% (SD = 2.4%), the geophone at 90.7% (SD = 4.0%), and the VPU at 95.0% (SD = 2.4%). As an additional point of reference, we observed similar performances for each device when using a Linear SVM (Scikit-Learn, default parameters): 75.8%, 83.9%, 93.6%, and 96.4% for the accelerometer, traditional microphone, geophone, and VPU, respectively. We want to explicitly note that these accuracy numbers are not necessarily representative of the maximum performance or the total number of the types of events that can be

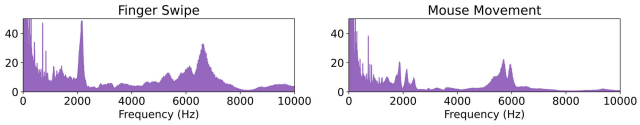

Figure 6: The frequency signature for two of the more subtle events, the swipe gesture (left) and mouse movement (right) when captured by the VPU.

achieved by each sensor, but are used to compare the relative performance across sensors for this task and illustrate the types of events each sensor is capable of recognizing.

For the accelerometer, we observed reasonable performance in classifying impulse-based events, like placing objects on a surface, but a significant weakness in more subtle events, like moving the mouse or swiping a finger, which incurred substantial confusion with the "nothing" class. We observed a similar but less pronounced effect for the regular microphone, which in a quiet environment, improved upon the classification of impulse-based events but still struggled with capturing signals from subtle events. The geophone had improved overall performance compared to the prior two sensors, but the "nothing" vs. "mouse" comparison remained the single largest source of error. The VPU overall had the best performance, with the best individual performance in 6 out of the nine classes. Its largest source of error was the item placement classes, such as "bottle", "keys", and "phone".

When we use the VPU to examine the frequency components of the "mouse" and "swipe" class, we see in Figure 6 that mouse and swipe movements have significant frequency components around 2.5kHz and above 5kHz. The significance of these frequency components offers an explanation for why the classification performance in these classes is reduced for the accelerometer/geophone, as they do not have the bandwidth to capture these higher frequency components. For the microphone, even though it can capture these frequencies, environmental noise (such as white noise from the building AC) even in a relatively quiet location, overlaps with these frequency ranges making it more difficult to isolate and thus classify these kinds of events. As a remedy, prior works, such as Scratch Input [24], use a high-pass filter at 3kHz to remove ambient and environmental noise. Reconducting the evaluation with a high-pass filter for the regular microphone, we observe a significant improvement in classification performance of the impulse-based classes, such as "keyboard" and "keys", but at a significant cost to "mouse" and "swipe" and an overall decrease in performance to 73.0%. For clarification, prior works such as SurfaceVibe [50] (geophone) and Scratch Input [24] (microphone) use fingernails for their tap and swipe classification; in this evaluation, we do not use fingernails.

Overall, through these evaluations, the VPU has shown a unique combination of elements from the previous sensors – small MEMS fabrication (accelerometer, microphone), robustness to environmental noise (accelerometer, geophone), wide frequency response (microphone), long-range sensing (geophone, microphone). Most importantly, VPUs capture signals in what seems to be an important region of frequencies that are present in many surface-based events, but cannot be captured by accelerometers and geophones and can be buried in ambient noise when captured with microphones. We quantify the importance of these frequencies in the next section.

4 SYSTEM DESIGN

The previous section demonstrated the unique qualities and capabilities of VPUs that offer advantages over accelerometers, geophones, and microphones for surface-bound sensing. Having selected the VPU as the basis for SAWSense, in this section, we detail the hardware implementation and include various design considerations for those who wish to utilize this sensing approach. We then detail the design of our software pipeline for real-time collection and classification of events, which includes an information power-based approach to craft an optimized featurization schema for the VPU, demonstrating improved classification performance over a baseline FFT approach. We now describe each in further detail.

4.1 Hardware Details

The Sonion Voice Pick Up (VPU) [11] consists of a mass-spring diaphragm mounted on top of the sound port of a TDK InvenSense ICS-40619 MEMS microphone and then hermetically sealed with a cover. A vent hole to relieve pressure changes during the reflow soldering process is sealed with epoxy per manufacturer specifications, making the VPU "virtually insensitive to acoustic signals" [11]. Movement causes the mass/spring diaphragm to oscillate, creating (sound) pressure which is measured by the MEMS microphone. Both the digital output Pulse Density Modulation (PDM) and analog differential output versions of the device were evaluated, and no significant differences were found.

In this work, we utilize the PDM implementation and place all components on a single 12mm X 15mm PCB for easier soldering (the VPU itself is 2.65mm X 3.5mm), as seen in Figure 7. We use a MiniDSP MCHStreamer to perform PDM conversion to a 48kHz 16-bit datastream over USB Audio Class (UAC), which allows for low-latency data transfer and is supported by most operating systems and useful libraries (e.g., PyAudio [52]). Since our repurposing of the VPU significantly differs from its initial use cases (contact with in-ear soft tissue for voice pickup), we performed some initial testing on whether factors such as contact angle or weight on the sensor affected the signal output. We note that there is a rectangular patch on the flat surface of the VPU where the device is sensitive and can capture signals. This patch needs to be in contact with the surface of interest. The SAWs cannot be picked up if the PCB or housing of the sensor is in contact with the surface but not the patch itself. For tables and other flat surfaces, a simple 3D-printed housing was used to provide strain relief for the cable and orient the VPU towards

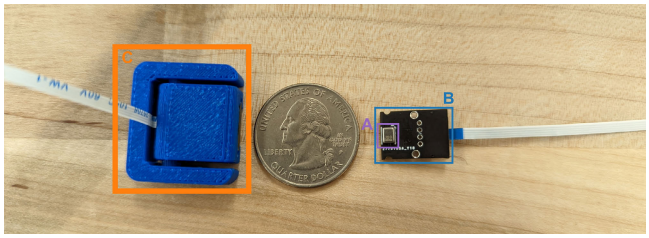

Figure 7: The size of the VPU (A), sensor PCB (B), and 3D printed housing (C) relative to a quarter.

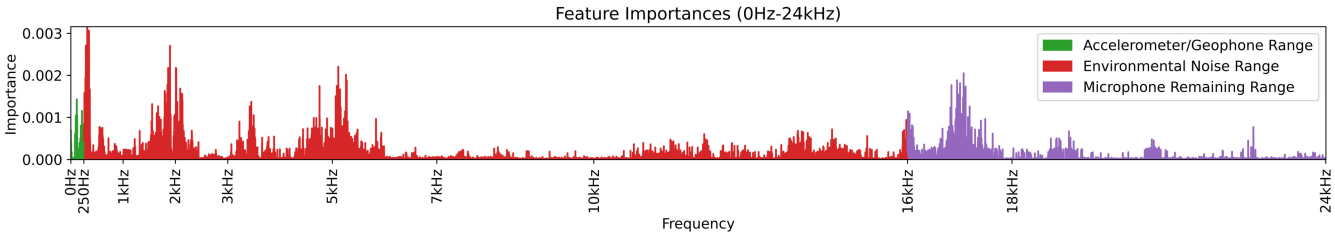

Figure 8: The feature importances across frequencies for the VPU. The frequency range in green denotes the frequencies an accelerometer or geophone can sense (0-250Hz). The frequency range in red denotes the range of typical environmental sounds that affect microphone performance but not VPU performance. The frequency range in purple denotes the remaining frequencies a microphone can sense. Overall, the range that accelerometers or geophones cannot capture and microphones incur noise (250Hz-16kHz) represents 75.4% of the total information power available, providing a significant advantage for VPUs in classification tasks.

the surface. No additional surface preparation was required. For odd geometries and orientations, we affixed the VPU in place with tape without any other modifications, as seen in the Video Figure. We note that the 3D-printed housing had no observable effects on noise isolation over the tape. Figure 7 shows the size of the VPU sensor and its 3D printed housing relative to a US quarter dollar.

4.2 Feature Importance and Validation

While the SAW signals collected in the previous section contain frequency components in what appears to be an important range for sensing, we can quantify the importance of each frequency band and determine their contribution to classification performance. While there are many ways to quantify feature importance, Gini Impurity offers a closely related metric to actual classification performance [46]. Thus, we use a Random Forest classifier to generate Gini Impurity for each frequency bin using computed FFTs features from the previous section. Figure 8 shows the importances from 0-24kHz. We observe that frequencies higher than 250Hz (above accelerometer/geophone capabilities) contribute 96.9% of the total information power available. Furthermore, when using a microphone, typical environmental sounds (e.g., speech up to 8kHz, music up to 16kHz) can introduce noise and bury frequencies that represent 78.3% of the total information power. Ultimately, the distribution of feature importances highlights the value of capturing higher frequency signals without also capturing overlapping environmental noises.

We used this Gini analysis to inform the design of an optimized featurization schema for SAW signals. While FFTs offer a reasonable baseline for offline evaluation, the large number of features may slow down classification performance. Ideally, we find a way to represent the most valuable frequency bands and capture the greatest amount of information while using the fewest number of features. Given the distribution of the feature importances, Mel-Frequency Cepstral Coefficients (MFCC) appear to be an appropriate choice to featurize SAW signals. However, the traditional mel-scale does not allocate features in the frequency ranges that would maximize the capture of important frequency bands. Fortunately, the Python library, librosa [44], can efficiently implement a highly-optimized custom MFCC (cMFCC) to allocate 128 mel filter banks from 0Hz to 18kHz, which represents a frequency range containing 94.5% of the feature importances. We create a clip-length cMFCC by summing all components over time. Overall, we reduce the number of features from 24001 to 128, which should also improve model robustness and performance.

To confirm that A) this approach matches or improves upon the performance of FFTs and B) reduces the prediction time, we rerun the cross-validation and find a 98.4% (SD = 1.1%) classification performance, compared to 94.8% with FFTs. The prediction time improved from 30ms with FFTs to 20ms with cMFCCs using an Intel i7-1185G7. For completeness, using a linear SVM on the Intel i7 found similar classification performance: 95.1% with FFT and 99.4% with cMFCC. Given that a linear SVM uses all of the features (and does not perform feature selection as part of the algorithm such as with the Random Forest), improvement in prediction time was more pronounced, from 9ms to 0.5ms, suggesting reasonable performance can be achieved in embedded and compute-constrained environments. In the following section, SAWSense is evaluated using this pipeline across applications in two distinct HCI domains.

5 SYSTEM EVALUATION

In this section, we evaluate SAWSense's ability to support applications across two distinct HCI domains, gestural input and activity recognition, to illustrate the breadth and versatility of SAWs for recognizing everyday surface-bound events. We first evaluate hand-to-surface trackpad-style gestures (e.g., taps, swipes, and flings) on a desk and demonstrate that SAWSense has the sensitivity to sense subtle gestures (such as finger flings) while also having the bandwidth to determine whether the finger pad or the fingernail was used to perform the gesture, effectively creating two "modes" for each gesture. We then explore the variety of SAWs generated by household appliances and cooking-related events and evaluate SAWSense's performance in activity recognition tasks in the kitchen. For both sets of evaluations, we present the results using a linear SVM, a Random Forest, and a 6-layer Multi-Layer Perceptron, representing three "tiers" of machine learning complexity. For consistency, all parameters except for the number of Random Forest estimators remain SciKit-Learn default. The number of estimators was tuned using orders of ten (N=10,100,1000,10000) and found N=1000 provided consistent performance for all experiments with reasonable training time. N=10000 provided marginally better performance, but at the cost of significantly increased training time.

Figure 9: This figure shows the relative placement of the sensor to where the gestures were performed and the desk used for this evaluation.

Across these two domains, we found robust performance, with each reporting greater than 97% accuracy. We now describe each evaluation in further detail. The following evaluations were conducted in accordance with our institution's Institutional Review Board (IRB).

5.1 Hand Gestures on Desk

In this evaluation, we measure our system's ability to discern subtle events through three common trackpad-style gestures (taps, swipes, and flings). While prior works have explored these gestures using the fingernail, we also record the gestures done using the fleshy pad of the finger, resulting in six total gestures. We envision the fingerpad vs. fingernail distinction for each base gesture could offer alternate input "modes" for interfaces; for example, a fingerpad tap could trigger a left click while a fingernail tap could trigger a right click. Figure 10 provides a depiction of these gestures and their average frequency signature. Additionally, since these gestures can be person-dependent, where finger/nail texture, location/force applied, and duration of gesture can influence the signal, we recruited participants and evaluate both per-user and generalized accuracy. We also evaluate the effectiveness of augmentation techniques with SAWs to generalize surface touch gesture performance.

5.1.1 Procedure. We recruited 10 participants (four female, six male – three had long natural nails, and one had acrylic nail extensions) and asked them to perform multiple data collection rounds consisting of six gestures ten times in a random order to introduce variety and prevent capturing nearly-identical events. A "nothing" class was also collected as part of the round, where no gesture was being performed. To keep the length of the study reasonable for participants, we asked them to repeat each round 5 times, which took approximately an hour and resulted in 350 total instances per participant (10 instances x 5 rounds x 7 classes). The full set of gestures was demonstrated once to the participant, but each participant was permitted to perform the gesture by what felt natural to them, allowing for variations in gesture location, force, and duration. It should be noted that participants with longer nails performed the finger pad gestures with their fingers parallel to the table to avoid contacting the table with their fingernails. Conversely, participants with short fingernails opted to perform the finger pad gestures with their fingers at higher angles while not touching the fingernails to the table. All participants had similar figure orientations for the three nail-based gestures. The study was conducted in an open office/lab environment using a wooden table (Ikea Bjursta). The sensor was removed and repositioned at least 60cm away for each participant to add variety and realism to the dataset. The dataset was collected over the course of a week, during which the table was moved, used by other occupants of the room, etc. We note that other persons were in the room, working at adjacent desks, walking, speaking, etc., and we did not control for typical office noise and activities.

5.1.2 Per-User Accuracy. We first evaluate our per-user accuracy by performing a leave-one-round-out cross-validation, where the classifier is trained on four rounds and tested on a 5th. We repeat this process for all combinations and compute the average result per participant. Across all participants, we found a mean per-user accuracy of 97.2% (SD=1.3%) when using an MLP Classifier (SciKit-Learn, relu activation, adam solver, 6 layers: 1024, 512, 256, 128, 64, 32). The mean confusion matrix can be found in Figure 11. We observe that nail-based gestures had higher classification accuracy than their finger-based counterparts, but all gestures had 93% accuracy or better. A t-SNE plot in Figure 11 visualizes a single participant's data, showing that the classes have minimal overlap except for finger fling and nail fling. This overlap correlates to finger fling vs nail fling confusion, which leads to the largest source of error. For completeness, we also perform the cross-validation using a linear SVM (default parameters) and a Random Forest (1000 estimators, all other parameters default). The results can be seen in Table 1 which show robust performance even with relatively simple ML approaches, suggesting that even greater performance can be achieved with more advanced or state-of-the-art models.

5.1.3 Generalized Accuracy. While the per-user results are promising, a more challenging evaluation is determining how effective our system can perform gesture recognition tasks when A) the location of the sensor is not fixed relative to the surface or interaction area, and B) individuals perform the gestures with different styles. We evaluate this "generalized" accuracy by training our system on all but one participant and test on the remaining participant. This allows us to roughly evaluate how our system would perform for a "new" user. We repeat this process for all combinations and report the average results. Overall, we found a mean accuracy of 95.2% (SD=1.0%) using the MLP classifier across all participants. The mean confusion matrix can be found in Figure 11. The results using other classifiers can be seen in Table 1.

We suspect the lower performance can be attributed to the size of our dataset relative to the number of factors that influence how the gesture is performed; typical sound recognition systems train off thousands of labeled instances per class in order to better represent the true diversity within that class [57, 58]. Taps gestures are performed rather consistently across users, hence their relatively high accuracy. However, we observed significant variations in how each participant performed the swipe and fling gestures

Table 1: The classification performance for gesture recognition in the per-user, across-user, and across-users with augmentation conditions across various machine learning classification approaches.

SciKit-Learn Classifier	Per-User	Across-User	Across-User w/ Augmentation
Linear SVM (default parameters)	**97.5% (SD=1.4)**	94.3% (SD=0.6%)	89.6% (SD=0.3%)
Random Forest (1000 estimators)	94.7% (SD=3.3%)	92.5% (SD=1.4%)	95.4% (SD=0.2%)
MLP (*relu*, *adam*, 6 layers: {1024, 512, 256, 128, 64, 32})	97.2% (SD=1.3%)	**95.2% (SD=1.0%)**	**97.6 (SD=0.3%)**

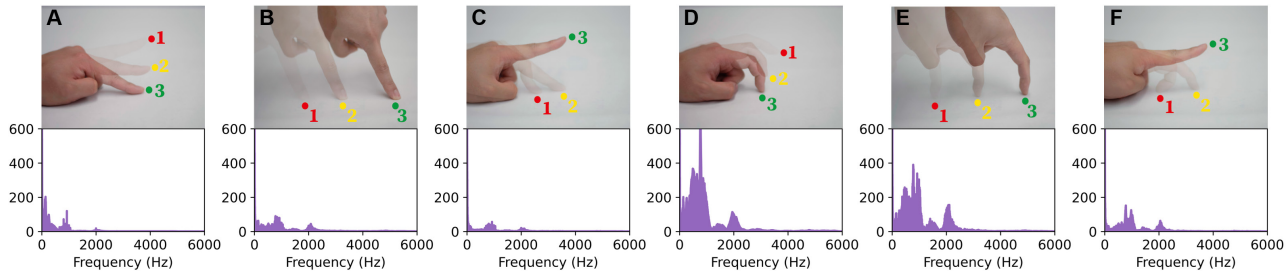

Figure 10: The six trackpad-style gestures with their motions are depicted in the top row and their frequency signatures in the bottom row: finger tap (A), finger swipe (B), finger fling (C), nail tap (D), nail fling (E), nail swipe (F).

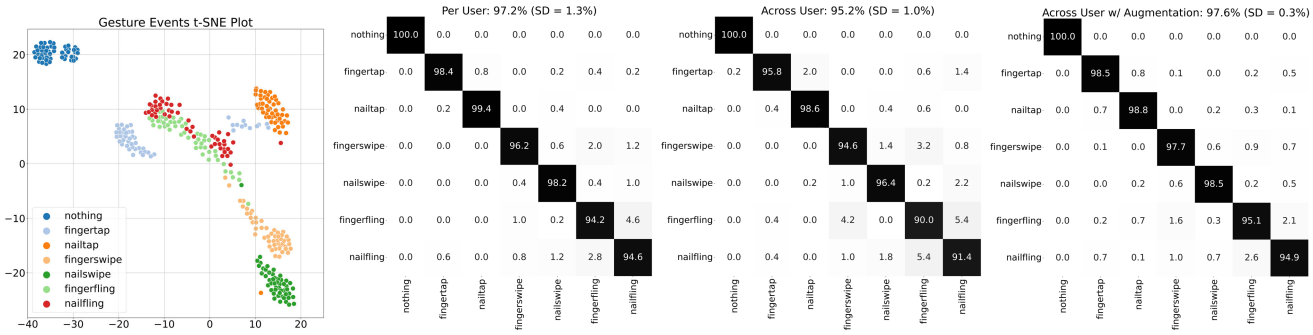

Figure 11: The t-SNE plot for one participant visualizing the different gesture classes and the confusion matrices using the MLP classifier for the per-user, across-users, and across-users with synthetically augmented SAW training data conditions.

(e.g., pressure, location, duration, fingernail length, presence of hand lotion/moisturizer on the skin). Significantly more participants would be needed to better represent all the different ways people can perform the same gesture. We observed that ~47% of the total error was attributed to finger vs. nail errors (i.e., getting the base gesture right but the fingertip type used wrong). This may be in part due to participants with short nails also using their finger pad as part of the gesture, and those with longer nails may have lightly brushed their fingernails during finger pad gestures.

5.1.4 Use of Synthetic Data with SAWs. To improve generalized accuracy across participants, we explored the use of traditional audio augmentation approaches to improve model robustness to real-world user conditions. We generate additional synthetic data from our real collected data, which can simulate a variety of different environmental noise, users' touch duration, force, range, and surface textures to ensure our system can maintain high accuracy for previously unseen users and usage conditions. We utilize the Python package, Audiomentations [34], to create augmented versions of our dataset as follows:

(1) AddGaussianSNR(min_snr_in_db=0, max_snr_in_db=50, p=.25) to simulate increased environmental noise.
(2) TimeStretch(min_rate=0.8, max_rate=1.25, leave_length_unchanged=True, p=.25) to simulate different gesture durations.
(3) PitchShift(min_semitones=-4, max_semitones=4, p=.25) to simulate different finger/nail conditions.
(4) Gain(min_gain_in_db=-12, max_gain_in_db=12, p=0.25) to simulate different gesture force.

Using our original dataset, we generate 20 instances from each collected instance, resulting in 7000 instances per participant. We featurize each instance using our cMFCC as before. We employ an MLP Classifier, which can take advantage of the larger augmented dataset, and perform a similar evaluation as before, where we train on all augmented instances of nine participants and test on all augmented instances of the 10th for all combinations, and report the averaged results. We observe an improved accuracy of 97.6% (SD = 0.3%) for all gestures, much closer to our per-user performance. The mean confusion matrix can be found in Figure 11. For completeness, Table 1 shows the performance using other ML approaches. We attribute the linear SVM's decreased performance with augmentation to the increase in the variety of the dataset, such that classes can no longer be cleanly separated with a straight line. Overall, we observe that traditional audio augmentation approaches can improve model performance for SAW classification tasks and help quickly bootstrap small datasets.

5.2 Activity Recognition in the Kitchen

Acoustic activity recognition is most commonly associated with the classification of audible sounds captured by microphones. In this evaluation, we explore the presence of SAWs in a home kitchen environment and use SAWSense to perform activity recognition tasks on a kitchen counter, such as detecting the operation of appliances (e.g., blender, mixer), cooking-related events (e.g., chopping, peeling), and placement of kitchenware (e.g., forks, mugs).

Figure 12: This figure shows the relative placement of the sensor on the wood kitchen counter used for this evaluation.

5.2.1 Procedure. We identified a number of appliances (food processor, stand mixer, blender, air fryer, coffee grinder, water boiler, microwave), cooking actions (whisking, opening microwave door, peeling, chopping), and placement of objects (fork, bowl, mug) that are representative of typical events that happen on kitchen surfaces. For the appliance classes, if the device had speed settings, we selected the lowest speed and highest speed and recorded them as separate classes. For each class, a 1-second clip was captured while the device was operating or action was being performed, creating a single instance. 10 instances of each class were collected, forming a round. 10 rounds in total were collected, resulting in 100 total instances across 17 classes (which includes the "nothing" class). In this evaluation, we did not control for environmental noise such as speech or music, other occupants, and pets in the home, similar to real-world conditions. Since this study focuses on objects rather than users, no participants were recruited. Figure 13 shows an image of each class and its corresponding frequency signature.

5.2.2 Results. We find through a 10-round cross-validation that the average performance across 17 classes is 99.3% (SD = 0.7%) using a Random Forest classifier. We see among the appliances that there is virtually no confusion, as their frequency signatures are very distinct. In the remaining classes, there is confusion between classes that overlap; peeling and chopping both use a cutting board on the surface of the counter; the bowl and mug are both ceramic, roughly the same weight, and have similar frequency signatures. Despite other events occurring in the home during data collection, the nothing class was very consistent, with no confusion with other classes. Overall, this evaluation suggests SAWs offer a compelling approach for in-home activity recognition systems, given their robustness to sounds and speech. The t-SNE plots and confusion matrices can be found in Figure 14. We also observed robust performance across ML approaches, as seen in Table 2.

5.2.3 Performance Across Materials. To better understand how the classification performance of SAWSense can generalize across materials, the evaluation was reconducted over one month after the wooden kitchen counter evaluation using a metal-top kitchen island, seen in Figure 15, with the same objects and sensor hardware. Ten rounds resulting in 100 instances per activity were collected in the same fashion as on the wooden countertop. In performing a similar 10-round cross-validation, the average performance across 17 classes is 99.2% (SD = 0.4%) using a Random Forest classifier.

However, when using the wooden kitchen counter's data to train the metal kitchen island, and vice versa, the classification performance is relatively poor: 59.2% (SD = 9.6%). Upon examining the waveforms and features of the classes directly, it appears that while the spectral signature between classes appears similar, the metal kitchen island has significantly less attenuation resulting in frequency components with significantly greater amplitude, leading to poor performance when used as features for a model trained on wood (and vise versa). We also observed that the attenuation was not a fixed amount across frequencies, and there is a "frequency response" to each material.

Rather than collecting instances for every combination of event and surface material, one potential avenue to bridge this gap is to create a function that transforms the features of one material into "synthetic" features of another material. We created two functions (i.e., "wood2metal", "metal2wood") by computing the mean of features across all events and calculating the adjustment to transform the cMFCC features of one material into generated features of the other material. We perform an evaluation, where metal2wood features are used to train a model and tested on wood and wood2metal features are used to train a model and tested on metal. In this evaluation, we found restored performance, 98.0% (SD = 0.5%) using the Random Forest classifier, but only marginally improved performance using the MLP. While we did not evaluate these kitchen activities on other home surfaces, such as glass or granite, our earlier evaluation showed that subtle taps and swipes could launch SAWs on those materials and thus expect the much more energetic kitchen activities to also do the same. Additionally, while we did not evaluate on larger surfaces of the same material, since SAWs can travel relatively long distances and given how "loud" kitchen activity SAWs are compared to taps and swipes, we expect similar performance without the need for a new model for the same material. Overall, these results indicate a promising path forward for creating material invariant models. For example, it should be possible to collect a small amount of data on a new material to determine its transfer function (from one material to another) and then use training data from a well-studied material, to create a robust model for the new material, thus reducing the need to collect extensive training data for all surface types.

6 DISCUSSION

In this section, we provide a brief discussion of the results in the previous sections, explain known limitations to SAWSense's sensing approach, and present avenues for future work.

Table 2: The classification performance for gesture recognition in the wood, metal, and cross-material conditions across various machine learning classification approaches.

SciKit-Learn Classifier	Wood	Metal	Cross-Material
Linear SVM (default parameters)	**99.5% (SD=0.5%)**	**99.5% (SD=0.5%)**	88.8% (SD=0.2%)
Random Forest (1000 estimators)	99.3% (SD=0.7%)	99.2% (SD=0.4%)	**98.0% (SD=0.5%)**
MLP (*relu*, *adam*, 6 layers: {1024, 512, 256, 128, 64, 32})	95.7% (SD=4.3%)	95.3% (SD=5.8%)	66.2% (SD=1.7%)

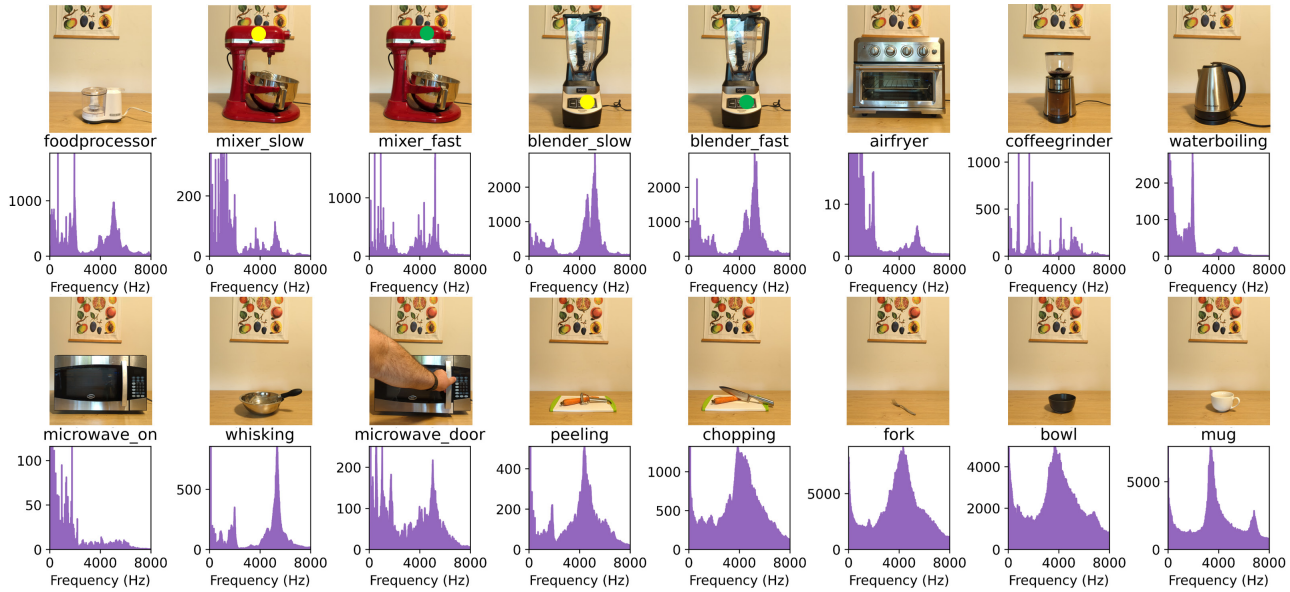

Figure 13: The sixteen different kitchen appliances, actions, or objects, and their frequency signatures. Please note the upper limit on the y-axis is adjusted in each class for visualization purposes.

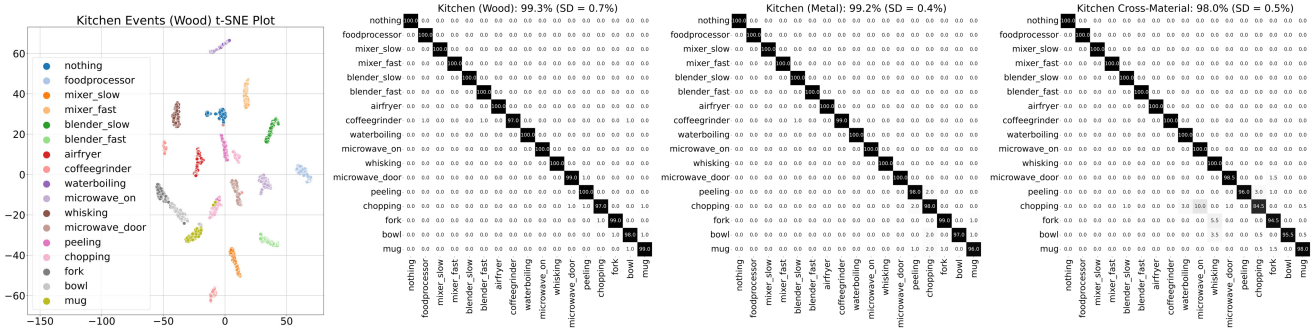

Figure 14: The confusion matrices show classification accuracy for the 17 different classes using the Random Forest classifier.

Figure 15: This figure shows the metal top kitchen island. The sensor was placed similarly to the wooden kitchen counter in the bottom left corner.

6.1 Classification Performance, Revisited

While SAWSense achieved reasonable performance in the previous sections, we note that A) the same featurization schema was used for all of the evaluations and B) the ML approaches are effective but by no means state-of-the-art. Our intention in using linear SVM, Random Forest, and Multi-Layer Perceptron was to show three different "tiers" of ML and how the performance is robust regardless of ML, demonstrating the utility and expressivity of the sensor itself without having to rely on advanced signal processing and ML approaches. For example, the linear SVM was able to achieve >88% performance in every task with only 128 features. We believe there remains significant headroom for future researchers and developers to optimize system performance.

More broadly speaking, while advanced approaches can be used to squeeze out maximum performance, these approaches often require very large training data sets, leading to large model sizes

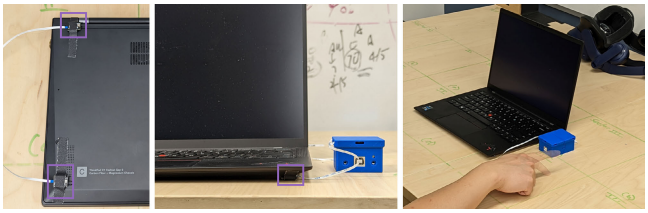

Figure 16: This figure shows how the two sensors were affixed to the two right feet of the laptop (left), how the laptop was placed on the desk with the two sensors in contact with the desk (middle), and where the gestures were performed relative to the laptop (right).

and expensive computing hardware and GPUs that increase the overall complexity and cost of the system as a whole. The simple ML approaches used in this work, highlighted in Future Work examples below, can be deployed on limited computing resources, such as microcontrollers (e.g., Arduino Nano 33) and microprocessors (e.g., Raspberry Pi), matching the overall spirit of SAWSense's low-cost, low-power sensing approach, which enables future avenues for widespread deployment and adoption.

6.2 Sensing Limitations

While SAWs are robust to external sounds and vibrations, similar to traditional acoustic methods, they are still susceptible to the "multiple sounds" problem, when multiple surface acoustic events are happening simultaneously. For traditional sound methods, multiple microphones can be used to perform sound source separation, such as with Independent Component Analysis (ICA), and each separated sound can be classified individually. We expect SAWs to be no different, but with only a single VPU, only one sound can be classified at a time.

One inherent limitation of SAW-based surface sensing approaches is that events can only be detected on the instrumented surface. For example, if an event happened on one surface (e.g., a kitchen counter) and there is no physical connection to another surface (e.g., a standalone kitchen island), a single VPU cannot capture events that occur on both surfaces (unless the event is very intense and can travel through the floor). This is a tradeoff with being robust to environmental noise. Conversely, since the VPUs can only pick up events on the surfaces they are connected to and cannot pick up sound from the environment, such as speech, they are innately privacy-preserving and offer a viable approach for always-on sensing in the home.

6.3 Future Work

6.3.1 Adding a Second Channel for Direction Information. We also explore adding a second sensor to see if the direction can be inferred as part of a recognized gesture event (e.g., swipe up vs. swipe down). We place two sensors on the two right feet of a ThinkPad laptop, securing them in place with tape without covering the VPU. The laptop is then placed on the same desk as before, where the VPUs are in contact with the surface of the desk. We collect the following gestures: fingertap towards the back of the laptop ("PgUp"), fingertap towards the front of the laptop ("PgDn"), fling towards the back of the laptop ("scroll_down"), and a fling towards the front of the laptop ("scroll_up"), and the "nothing" class. We collect 10 instances of each gesture, forming a round. We collect 10 rounds in total, resulting in 100 total instances per gesture. Since there are 2 channels, we compute the cMFCC for each channel and also include a single computed feature of which channel has the greater magnitude. Using a second channel, we could distinguish the gesture and its direction robustly, with 99.6% (SD = 0.8%) accuracy using a linear SVM. This technical demonstration suggests a low-cost and compelling method for extending the interaction area of mobile devices onto the surface of the table around them.

6.3.2 Support for Odd Geometries and Interaction with Printed Objects. While the previous sections explore the utility of SAW-based sensing on flat surfaces, electronics manufacturers and product designers often need to create sensing devices and user interfaces on non-planar geometries. However, manufacturing conformal capacitive and resistive sensing arrays on oddly shaped objects can be labor-intensive and/or prohibitively expensive. Surface Acoustic Waves will continue to propagate along an unbroken surface, meaning that the instrumentation point does not have to be on the same side as the area of interaction. A simple example of this is capturing interactions on the top of a table by taping a VPU to the underside of the table, as seen in the Video Figure. To demonstrate the effectiveness of SAWSense at enabling complex user interfaces with a single point of instrumentation on much more complex geometries, a children's dragon toy, which does not have a flat surface apart from the bottom of its feet and is not of homogeneous construction (i.e., the surface was not consistently of one material) has been created as shown in Figure 17.

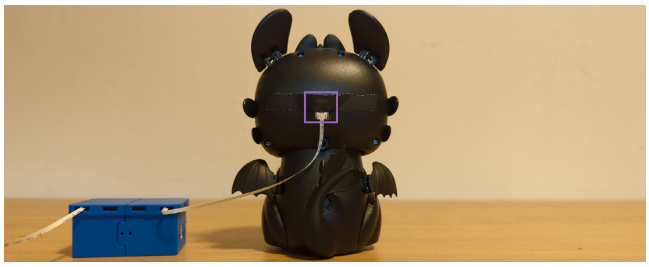

Figure 17: This figure shows how the sensor was affixed to the back of the head of the toy dragon.

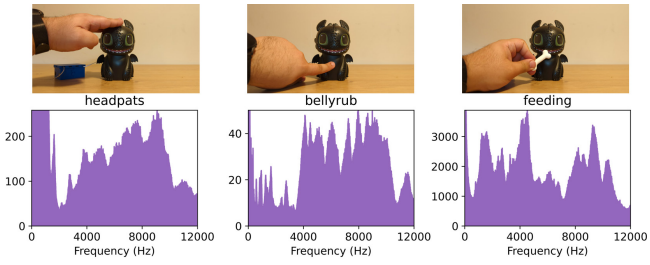

Figure 18: The three interactions with the toy dragon and their frequency signatures. Please note the upper limit on the y-axis is adjusted in each class for visualization purposes.

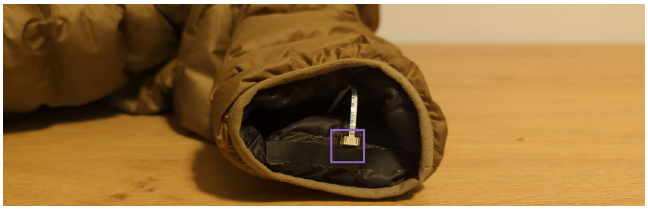

Figure 19: This figure shows how the sensor was affixed to the inside of the jacket sleeve near the cuff.

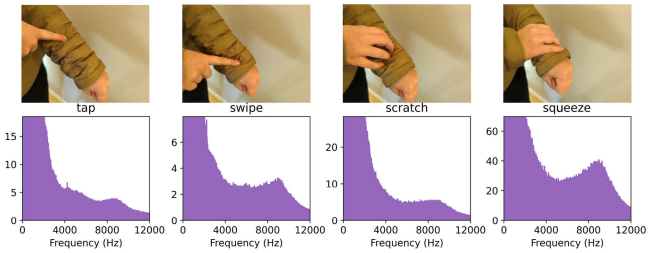

Figure 20: The four on-sleeve interactions and their frequency signatures. Please note the upper limit on the y-axis is adjusted in each class for visualization purposes.

In the spirit of playful interaction, we select two hand gestures, a head pat, and a belly rub. We also 3D printed a small bone to "feed" the dragon by tapping the bone against its mouth. A "nothing" class is also collected. For each class, a 1-second clip was captured while the action was being performed, creating a single instance. 10 rounds of 10 instances per class were collected, resulting in 100 instances per class. Figure 18 shows an image of each class and its corresponding frequency signature.

We perform a 10-round cross-validation, using the same pipeline as in the previous evaluations, and find a 100.0% (SD = 0.0%) classification accuracy with a Random Forest. Given how different the frequency signature of each class is, we do not find these results surprising. However, we do realize that a 48kHz sampled, 128-cMFCC, Random Forest pipeline is quite excessive for recognizing gestures on a children's toy, given the typical constraints (cost, power) associated with integrating electronics into toys.

Thus, since many children's toys contain microcontrollers, we also simulate the performance of the system running on a low-cost embedded device, such as a Nordic NRF52. For reference, the NRF52 series can hardware-decode up to 2 PDM streams at 16kHz 16-bit sampling. Thus, we modify our cMFCC to craft 20 features from 0Hz to 8kHz (Nyquist limit) and use a linear SVM to perform predictions, both of which can be performed in real-time locally on the NRF52. In our simulated performance on a microcontroller, we found a 99.7% (SD = 0.7%) classification accuracy. These results show promise that while SAWs can be used to perform more complex recognition tasks, the sensing approach remains accessible to simpler approaches on low-cost and low-power devices.

6.3.3 Support for Soft Surfaces and Fabrics. Finally, while the previous evaluations explored the presence and performance of SAWs on rigid surfaces, we also explored whether these signals can be captured on soft surfaces and used for event classification tasks. We first placed a sensor on some soft surfaces (e.g., leather chair seat, fabric couch arm) but quickly realized that while the stuffing in the furniture is soft, the surface material (leather, fabric) of the furniture is held taut, to which the signals appear only mildly attenuated compared to those on the rigid surfaces we described earlier. Inspired by other wearable gesture inputs, we instead look to explore soft fabrics that are not held taut, like a shirt or a light jacket. We selected a Patagonia Nanopuff jacket for this evaluation and present our results in two ways, one with the full pipeline and, similar to the toy example, one with a simulated embedded device.

We place the sensor on the inside of the left sleeve near the cuff and perform gestures with the right hand on the outside of the left sleeve near the cuff, in a similar interaction area to the Jacquard jacket [55]. The sensor has sufficient slack in the cable, so that movement in both arms remains unrestricted. However, since SAWSense's current implementation is over USB, the gestures are collected while seated and tethered to a computer. We adopt two of the same gestures from Jacquard, tap and swipe, and also add scratch and squeeze. We also collect the "nothing" class. Similar to previous evaluations, we collect a 1-second clip per instance, 10 instances per round, and 10 rounds total, resulting in 100 instances per class. Figure 20 shows an image of each class and its corresponding frequency signature. We note that unlike the "nothing" class in previous evaluations, the "nothing" class on the jacket has a number of nontrivial frequency components. Upon inspecting the raw waveforms, we see the presence of signals and confirm the origin of these signals to be caused by movement while wearing the jacket, such as the fabric of the jacket rubbing against itself.

We perform a 10-round cross-validation with the full pipeline, finding a 99.2% (SD = 1.3%) classification accuracy. Similar to the toy evaluation, we also evaluate the performance of the "embedded" pipeline with a 10-round cross-validation, finding 98.6% (SD = 1.0%) classification accuracy. While these results are promising, we wish to explicitly note that these results are meant to evaluate whether soft, non-taut fabrics are able to transmit SAWs and whether our system can classify them effectively. These results are not meant to represent the gesture recognition performance in active life, such as while walking or running.

7 CONCLUSION

This work has investigated methods for capturing Surface Acoustic Waves (SAWs) generated when a user or object touches or operates on a surface, using a Sonion Voice PickUp (VPU) bone sensor that has been repurposed as a high-bandwidth, high-isolation contact microphone. Experimental results show that SAWs exhibit unique propagation characteristics fundamentally different from sound or mechanical vibration, such as propagation along the surface-to-air boundary, and exhibit low attenuation compared to bulk waves that travel through an object. When captured with the VPU, the SAW signals can be detected across a wide range of material types, over relatively long distances on the order of meters, and are well-isolated from background noise which causes issues for traditional contact-based recognition systems.

With our signal processing and machine learning pipeline, SAWSense effectively detects surface gestures and surface-to-object interaction events with significantly higher accuracy than alternative

approaches. User evaluations show that our system can robustly recognize six surface gestures such as taps, swipes, and flings with >97% accuracy. For the first time, a data augmentation pipeline is demonstrated for Surface Acoustic Waves that shows the accuracy of cross-user gesture recognition can be significantly increased, reducing the need for extensive data collection for model training. Furthermore, object recognition accuracy for SAWSense is >99% without significant optimization, far outperforming alternative approaches while being easy to integrate into consumer electronics and deployed in activity detection applications.

To illustrate the utility of SAWs as a signal source for sensing surface interaction events, SAWSense was evaluated across two application domains, gesture input and ubiquitous activity recognition. On traditional flat surfaces, such as on a desk, SAWSense recognizes trackpad-style gestures and, with the aid of a second sensor, can also infer the direction of these gestures. In the home environment, SAWSense can perform activity recognition tasks and robustly classify 16 different cooking-related events, even in the presence of real-world noise, such as speech or music. Beyond flat surfaces, SAWSense can support playful interactions with an odd-geometry dragon toy and 3D-printed tangibles, even on a scaled-back pipeline that can run on microcontrollers. Finally, we used SAWSense to explore interactions on a jacket sleeve, showing SAWs are not limited to only rigid surfaces. In all of these evaluations, SAWSense achieved >97% classification accuracy. Ultimately, SAWSense demonstrates a low-cost and effective method for enabling computing systems to understand surface-based gestures, activities, and object interaction events that can be combined with other sensing modalities, or stand on its own, to enable a wide range of interactive and assistive technologies.

REFERENCES

[1] R.N. Aguilar and G.C.M. Meijer. 2002. Fast interface electronics for a resistive touch-screen. In *SENSORS, 2002 IEEE*, Vol. 2. 1360–1363 vol.2. https://doi.org/10.1109/ICSENS.2002.1037318

[2] AnalogDevices. 2019. Analog Devices ADMP 401. Website. Retrieved September 20, 2019 from https://www.analog.com/media/en/technical-documentation/obsolete-data-sheets/ADMP401.pdf.

[3] Juliette Artru, Thomas Farges, and Philippe Lognonné. 2004. Acoustic waves generated from seismic surface waves: propagation properties determined from Doppler sounding observations and normal-mode modelling. *Geophysical Journal International* 158, 3 (09 2004), 1067–1077. https://doi.org/10.1111/j.1365-246X.2004.02377.x arXiv:https://academic.oup.com/gji/article-pdf/158/3/1067/5985707/158-3-1067.pdf

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*. 892–900.

[5] Behringer. 2022. "Behringer UM2". Website. Retrieved April 7, 2022 from https://www.behringer.com/product.html?modelCode=P0AVV.

[6] Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon. 2018. *Sound Analysis in Smart Cities*. Springer International Publishing, Cham, 373–397. https://doi.org/10.1007/978-3-319-63450-0_13

[7] Hrvoje Benko, Andrew D. Wilson, and Federico Zannier. 2014. Dyadic Projected Spatial Augmented Reality. 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, 645–655. https://doi.org/10.1145/2642918.2647402

[8] Andreas Braun, Stefan Krepp, and Arjan Kuijper. 2015. Acoustic Tracking of Hand Activities on Surfaces. In *Proceedings of the 2nd International Workshop on Sensor-Based Activity Recognition and Interaction* (Rostock, Germany) *(iWOAR '15)*. Association for Computing Machinery, New York, NY, USA, Article 9, 5 pages. https://doi.org/10.1145/2790044.2790052

[9] Pia Breuer, Christian Eckes, and Stefan Müller. 2007. Hand gesture recognition with a novel IR time-of-flight range camera–a pilot study. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*. Springer, 247–260.

[10] Colin Campbell. 1989. 1 - Introduction. In *Surface Acoustic Wave Devices and their Signal Processing Applications*, Colin Campbell (Ed.). Academic Press, 1–7. https://doi.org/10.1016/B978-0-12-157345-4.50005-9

[11] Paul Clemens. 2018. Sonion Voice Pick up (VPU) sensor - TDK. https://invensense.tdk.com/wp-content/uploads/2018/10/Sonion-Voice-Pick-Up-VPU-Sensor-Paul-Clemens.pdf

[12] Analog Devices. 2022. "ADXL1001 Datasheet". Website. Retrieved April 7, 2022 from https://www.analog.com/en/products/adxl1001.html.

[13] Paul Dietz and Darren Leigh. 2001. DiamondTouch: A Multi-User Touch Technology. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (Orlando, Florida) *(UIST '01)*. Association for Computing Machinery, New York, NY, USA, 219–226. https://doi.org/10.1145/502348.502389

[14] Dodotronic. 2019. Ultramic384K. Website. Retrieved September 20, 2019 from https://www.dodotronic.com/ultramic384k/.

[15] Rick Downs. 2005. Using resistive touch screens for human/machine interface. *Analog Applications Journal Q* 3 (2005), 5–10.

[16] Bill Drafts. 2000. Acoustic Wave Technology Sensors. https://www.fierceelectronics.com/components/acoustic-wave-technology-sensors

[17] Benjamin Elizalde, Rohan Badlani, Ankit Shah, Anurag Kumar, and Bhiksha Raj. 2018. Nels-never-ending learner of sounds. *arXiv preprint arXiv:1801.05544* (2018).

[18] Kees Faber and Peter W. Maxwell. 2005. *Geophone spurious frequency: What is it and how does it affect seismic data?* 79–80. https://doi.org/10.1190/1.1826773 arXiv:https://library.seg.org/doi/pdf/10.1190/1.1826773

[19] Thomas Franke, Adam R Abate, David A Weitz, and Achim Wixforth. 2009. Surface acoustic wave (SAW) directed droplet flow in microfluidics for PDMS devices. *Lab on a Chip* 9, 18 (2009), 2625–2627.

[20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[21] Mayank Goel, Brendan Lee, Md. Tanvir Islam Aumi, Shwetak Patel, Gaetano Borriello, Stacie Hibino, and Bo Begole. 2014. SurfaceLink: Using Inertial and Acoustic Sensing to Enable Multi-Device Interaction on a Surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1387–1396. https://doi.org/10.1145/2556288.2557120

[22] Jun Gong, Aakar Gupta, and Hrvoje Benko. 2020. *Acustico: Surface Tap Detection and Localization Using Wrist-Based Acoustic TDOA Sensing*. Association for Computing Machinery, New York, NY, USA, 406–419. https://doi.org/10.1145/3379337.3415901

[23] Tobias Grosse-Puppendahl, Christian Holz, Gabe Cohn, Raphael Wimmer, Oskar Bechtold, Steve Hodges, Matthew S. Reynolds, and Joshua R. Smith. 2017. Finding Common Ground: A Survey of Capacitive Sensing in Human-Computer Interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3293–3315. https://doi.org/10.1145/3025453.3025808

[24] Chris Harrison and Scott E. Hudson. 2008. Scratch Input: Creating Large, Inexpensive, Unpowered and Mobile Finger Input Surfaces. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) *(UIST '08)*. Association for Computing Machinery, New York, NY, USA, 205–208. https://doi.org/10.1145/1449715.1449747

[25] Chris Harrison, Julia Schwarz, and Scott E. Hudson. 2011. TapSense: Enhancing Finger Interaction on Touch Surfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 627–636. https://doi.org/10.1145/2047196.2047279

[26] Chris Harrison, Robert Xiao, and Scott Hudson. 2012. *Acoustic Barcodes: Passive, Durable and Inexpensive Notched Identification Tags*. Association for Computing Machinery, New York, NY, USA, 563–568. https://doi.org/10.1145/2380116.2380187

[27] Alexander G. Hauptmann and Paul McAvinney. 1993. Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies* 38, 2 (1993), 231–249. https://doi.org/10.1006/imms.1993.1011

[28] Yan He, Hanyan Zhang, Edwin Yang, and Song Fang. 2020. Virtual Step PIN Pad: Towards Foot-input Authentication Using Geophones. In *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. 649–657. https://doi.org/10.1109/MASS50613.2020.00084

[29] David Holman, Nicholas Fellion, and Roel Vertegaal. 2014. Sensing Touch Using Resistive Graphs. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 195–198. https://doi.org/10.1145/2598510.2598552

[30] Yasha Iravantchi, Karan Ahuja, Mayank Goel, Chris Harrison, and Alanson Sample. 2021. PrivacyMic: Utilizing Inaudible Frequencies for Privacy Preserving Daily Activity Recognition. 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 198, 13 pages. https://doi.org/10.1145/3411764.3445169

[31] Hiroshi Ishii, Craig Wisneski, Julian Orbanes, Ben Chun, and Joe Paradiso. 1999. PingPongPlus: Design of an Athletic-Tangible Interface for Computer-Supported Cooperative Play. 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, 394–401. https://doi.org/10.1145/302979.303115

[32] Naoya Isoyama, Tsutomu Terada, and Masahiko Tsukamoto. 2014. An Interactive System for Recognizing User Actions on a Surface Using Accelerometers. In *Proceedings of the 5th Augmented Human International Conference* (Kobe, Japan) *(AH '14)*. Association for Computing Machinery, New York, NY, USA, Article 57, 2 pages. https://doi.org/10.1145/2582051.2582108

[33] Brett Jones, Rajinder Sodhi, Michael Murdock, Ravish Mehra, Hrvoje Benko, Andrew Wilson, Eyal Ofek, Blair MacIntyre, Nikunj Raghuvanshi, and Lior Shapira. 2014. RoomAlive: Magical Experiences Enabled by Scalable, Adaptive Projector-Camera Units. 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, 637–644. https://doi.org/10.1145/2642918.2647383

[34] Iver Jordal, Araik Tamazian, Emmanouil Theofanis Chourdakis, Céline Angonin, askskro, Nikolay Karpov, Omer Sarioglu, kvilouras, Enis Berk Çoban, Florian Mirus, Jeong-Yoon Lee, Kwanghee Choi, MarvinLvn, SolomidHero, and Tanel Alumäe. 2022. *iver56/audiomentations: v0.24.0*. https://doi.org/10.5281/zenodo.6367011

[35] Shrinivas G Joshi. 1991. Surface-acoustic-wave (SAW) flow sensor. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 38, 2 (1991), 148–154.

[36] KeySight. 2019. Agilent 33521A Function Generator. Website. Retrieved September 20, 2019 from https://www.keysight.com/en/pd-1871159-pn-33521A/function-arbitrary-waveform-generator-30-mhz?&cc=US&lc=eng.

[37] Hyosu Kim, Anish Byanjankar, Yunxin Liu, Yuanchao Shu, and Insik Shin. 2018. UbiTap: Leveraging Acoustic Dispersion for Ubiquitous Touch Interface on Solid Surfaces. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems* (Shenzhen, China) *(SenSys '18)*. Association for Computing Machinery, New York, NY, USA, 211–223. https://doi.org/10.1145/3274783.3274848

[38] Eric Larson, Gabe Cohn, Sidhant Gupta, Xiaofeng Ren, Beverly Harrison, Dieter Fox, and Shwetak Patel. 2011. HeatWave: Thermal Imaging for Surface User Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2565–2574. https://doi.org/10.1145/1978942.1979317

[39] Ching-Hua Lee, Bhaskar D Rao, and Harinath Garudadri. 2018. Bone-conduction sensor assisted noise estimation for improved speech enhancement. In *Interspeech*, Vol. 2018. NIH Public Access, 1180.

[40] Hanchuan Li, Can Ye, and Alanson P. Sample. 2015. IDSense: A Human Object Interaction Detection System Based on Passive UHF RFID. 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, 2555–2564. https://doi.org/10.1145/2702123.2702178

[41] Pedro Lopes, Ricardo Jota, and Joaquim A. Jorge. 2011. Augmenting Touch Interaction through Acoustic Sensing. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (Kobe, Japan) *(ITS '11)*. Association for Computing Machinery, New York, NY, USA, 53–56. https://doi.org/10.1145/2076354.2076364

[42] LSMDS1. 2022. "LSMDS1 Fast Library". Website. Retrieved April 7, 2022 from https://github.com/FemmeVerbeek/Arduino_LSM9DS1.

[43] Ben Maman and Amit Bermano. 2022. TypeNet: Towards Camera Enabled Touch Typing on Flat Surfaces through Self-Refinement. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 567–576. https://doi.org/10.1109/WACV51458.2022.00064

[44] Brian McFee, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Dan Ellis, Jack Mason, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, viktorandreevichmorozov, Keunwoo Choi, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Adam Weiss, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Matt Vollrath, Taewoon Kim, and Thassilo. 2022. *librosa/librosa: 0.9.1*. https://doi.org/10.5281/zenodo.6097378

[45] 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

[46] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics* 10, 1 (2009), 213.

[47] Jon Moeller and Andruid Kerne. 2012. ZeroTouch: An Optical Multi-Touch and Free-Air Interaction Architecture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2165–2174. https://doi.org/10.1145/2207676.2208368

[48] Takaaki Nara, Masaya Takasaki, Taro Maeda, Toshiro Higuchi, Shigeru Ando, and Susumu Tachi. 2001. Surface acoustic wave tactile display. *IEEE Computer Graphics and Applications* 21, 6 (2001), 56–63.

[49] Makoto Ono, Buntarou Shizuki, and Jiro Tanaka. 2013. Touch & Activate: Adding Interactivity to Existing Objects Using Active Acoustic Sensing. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 31–40. https://doi.org/10.1145/2501988.2501989

[50] Shijia Pan, Ceferino Gabriel Ramirez, Mostafa Mirshekari, Jonathon Fagert, Albert Jin Chung, Chih Chi Hu, John Paul Shen, Hae Young Noh, and Pei Zhang. 2017. SurfaceVibe: Vibration-Based Tap & Swipe Tracking on Ubiquitous Surfaces *(IPSN '17)*. Association for Computing Machinery, New York, NY, USA, 197–208. https://doi.org/10.1145/3055031.3055077

[51] J. A. Paradiso, K. Hsiao, J. Strickon, J. Lifton, and A. Adler. 2000. Sensor systems for interactive surfaces. *IBM Systems Journal* 39, 3.4 (2000), 892–914. https://doi.org/10.1147/sj.393.0892

[52] Hubert Pham. 2017. "PyAudio". Website. Retrieved September 17, 2020 from https://people.csail.mit.edu/hubert/pyaudio/.

[53] Karol J Piczak. 2015. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*. IEEE, 1–6.

[54] Louis Pisha, Julian Warchall, Tamara Zubatiy, Sean Hamilton, Ching-Hua Lee, Ganz Chockalingam, Patrick P Mercier, Rajesh Gupta, Bhaskar D Rao, and Harinath Garudadri. 2019. A wearable, extensible, open-source platform for hearing healthcare research. *IEEE Access* 7 (2019), 162083–162101.

[55] Ivan Poupyrev, Nan-Wei Gong, Shiho Fukuhara, Mustafa Emre Karagozler, Carsten Schwesig, and Karen E. Robinson. 2016. Project Jacquard: Interactive Digital Textiles at Scale. 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, 4216–4227. https://doi.org/10.1145/2858036.2858176

[56] Mark Richardson, Matt Durasoff, and Robert Wang. 2020. *Decoding Surface Touch Typing from Hand-Tracking*. Association for Computing Machinery, New York, NY, USA, 686–696. https://doi.org/10.1145/3379337.3415816

[57] Justin Salamon and Juan Pablo Bello. 2015. Unsupervised feature learning for urban sound classification. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 171–175.

[58] Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters* 24, 3 (2017), 279–283.

[59] SciKit. 2020. "SciKit". Website. Retrieved September 17, 2020 from https://scikit-learn.org/stable/.

[60] Yilei Shi, Haimo Zhang, Jiashuo Cao, and Suranga Nanayakkara. 2020. VersaTouch: A Versatile Plug-and-Play System That Enables Touch Interactions on Everyday Passive Surfaces. In *Proceedings of the Augmented Humans International Conference* (Kaiserslautern, Germany) *(AHs '20)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/3384657.3384778

[61] SM-24. 2022. "SM-24 Geophone". Website. Retrieved April 7, 2022 from https://www.iongeo.com/wp-content/uploads/2020/07/SM-24-Geophone-Element-Product-Flyer.pdf.

[62] Knowles SPH1668LM4H. 2022. "Knowles SPH1668LM4H". Website. Retrieved April 7, 2022 from https://www.digikey.com/en/products/detail/knowles/SPH1668LM4H-1/5332441.

[63] Sai Ganesh Swaminathan, Scott Hudson, and Steve Hodges. 2020. Using Surface Acoustic Wave Devices for Self-powered Sensing & Interaction. In *SelfSustainableCHI Workshop, CHI 2020*. https://www.microsoft.com/en-us/research/publication/using-surface-acoustic-wave-devices-for-self-powered-sensing-interaction/

[64] UMA-8. 2022. "MiniDSP UMA-8". Website. Retrieved April 7, 2022 from https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array.

[65] Pierre Wellner. 1991. The DigitalDesk Calculator: Tangible Manipulation on a Desk Top Display. In *Proceedings of the 4th Annual ACM Symposium on User Interface Software and Technology* (Hilton Head, South Carolina, USA) *(UIST '91)*. Association for Computing Machinery, New York, NY, USA, 27–33. https://doi.org/10.1145/120782.120785

[66] Andrew Wilson, Hrvoje Benko, Shahram Izadi, and Otmar Hilliges. 2012. *Steerable Augmented Reality with the Beamatron*. Association for Computing Machinery, New York, NY, USA, 413–422. https://doi.org/10.1145/2380116.2380169

[67] Andrew D. Wilson and Hrvoje Benko. 2010. Combining Multiple Depth Cameras and Projectors for Interactions on, above and between Surfaces. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 273–282. https://doi.org/10.1145/1866029.1866073

[68] Robert Xiao, Teng Cao, Ning Guo, Jun Zhuo, Yang Zhang, and Chris Harrison. 2018. LumiWatch: On-Arm Projected Graphics and Touch Input. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173669

[69] Robert Xiao, Chris Harrison, and Scott E. Hudson. 2013. WorldKit: Rapid and Easy Creation of Ad-Hoc Interactive Applications on Everyday Surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 879–888. https://doi.org/10.1145/2470654.2466113

[70] Robert Xiao, Scott Hudson, and Chris Harrison. 2016. DIRECT: Making Touch Tracking on Ordinary Surfaces Practical with Hybrid Depth-Infrared Sensing. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces* (Niagara Falls, Ontario, Canada) *(ISS '16)*. Association for Computing Machinery, New York, NY, USA, 85–94. https://doi.org/10.1145/2992154.2992173

[71] Robert Xiao, Greg Lew, James Marsanico, Divya Hariharan, Scott Hudson, and Chris Harrison. 2014. Toffee: Enabling Ad Hoc, around-Device Interaction with Acoustic Time-of-Arrival Correlation. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices; Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 67–76. https://doi.org/10.1145/2628363.2628383

[72] Yang Zhang, Gierad Laput, and Chris Harrison. 2017. Electrick: Low-Cost Touch Sensing Using Electric Field Tomography. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3025453.3025842

[73] Yang Zhang, Chouchang (Jack) Yang, Scott E. Hudson, Chris Harrison, and Alanson Sample. 2018. Wall++: Room-Scale Interactive and Context-Aware Sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 273, 15 pages. https://doi.org/10.1145/3173574.3173847