ActSonic: Recognizing Everyday Activities from Inaudible Acoustic Wave Around the Body

SAIF MAHMUD*, Cornell University, USA
VINEET PARIKH, Cornell University, USA
QIKANG LIANG, Cornell University, USA
KE LI, Cornell University, USA
RUIDONG ZHANG, Cornell University, USA
ASHWIN AJIT, Cornell University, USA
VIPIN GUNDA, Cornell University, USA
DEVANSH AGARWAL, Cornell University, USA
FRANCOIS GUIMBRETIERE, Cornell University, USA
CHENG ZHANG, Cornell University, USA

We present ActSonic, an intelligent, low-power active acoustic sensing system integrated into eyeglasses that can recognize 27 different everyday activities (e.g., eating, drinking, toothbrushing) from inaudible acoustic waves around the body. It requires only a pair of miniature speakers and microphones mounted on each hinge of the eyeglasses to emit ultrasonic waves, creating an acoustic aura around the body. The acoustic signals are reflected based on the position and motion of various body parts, captured by the microphones, and analyzed by a customized self-supervised deep learning framework to infer the performed activities on a remote device such as a mobile phone or cloud server. ActSonic was evaluated in user studies with 19 participants across 19 households to track its efficacy in everyday activity recognition. Without requiring any training data from new users (leave-one-participant-out evaluation), ActSonic detected 27 activities, achieving an average F1-score of 86.6% in fully unconstrained scenarios and 93.4% in prompted settings at participants' homes.

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

Additional Key Words and Phrases: Acoustic Sensing; Activity Recognition, Self-supervised Leaning

ACM Reference Format:
Saif Mahmud, Vineet Parikh, Qikang Liang, Ke Li, Ruidong Zhang, Ashwin Ajit, Vipin Gunda, Devansh Agarwal, Francois Guimbretiere, and Cheng Zhang. 2024. ActSonic: Recognizing Everyday Activities from Inaudible Acoustic Wave Around the Body. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 183 (December 2024), 32 pages. https://doi.org/10.1145/3699752

*Corresponding Author

Authors' Contact Information: Saif Mahmud, Cornell University, Ithaca, NY, USA, sm2446@cornell.edu; Vineet Parikh, Cornell University, Ithaca, NY, USA, vap43@cornell.edu; Qikang Liang, Cornell University, Ithaca, NY, USA, ql75@cornell.edu; Ke Li, Cornell University, Ithaca, NY, USA, kl975@cornell.edu; Ruidong Zhang, Cornell University, Ithaca, NY, USA, rz379@cornell.edu; Ashwin Ajit, Cornell University, Ithaca, NY, USA, aa794@cornell.edu; Vipin Gunda, Cornell University, Ithaca, NY, USA, vg245@cornell.edu; Devansh Agarwal, Cornell University, Ithaca, NY, USA, da398@cornell.edu; Francois Guimbretiere, Cornell University, Ithaca, NY, USA, francois@cs.cornell.edu; Cheng Zhang, Cornell University, Ithaca, NY, USA, chengzhang@cornell.edu.

ACM 2474-9567/2024/12-ART183
https://doi.org/10.1145/3699752

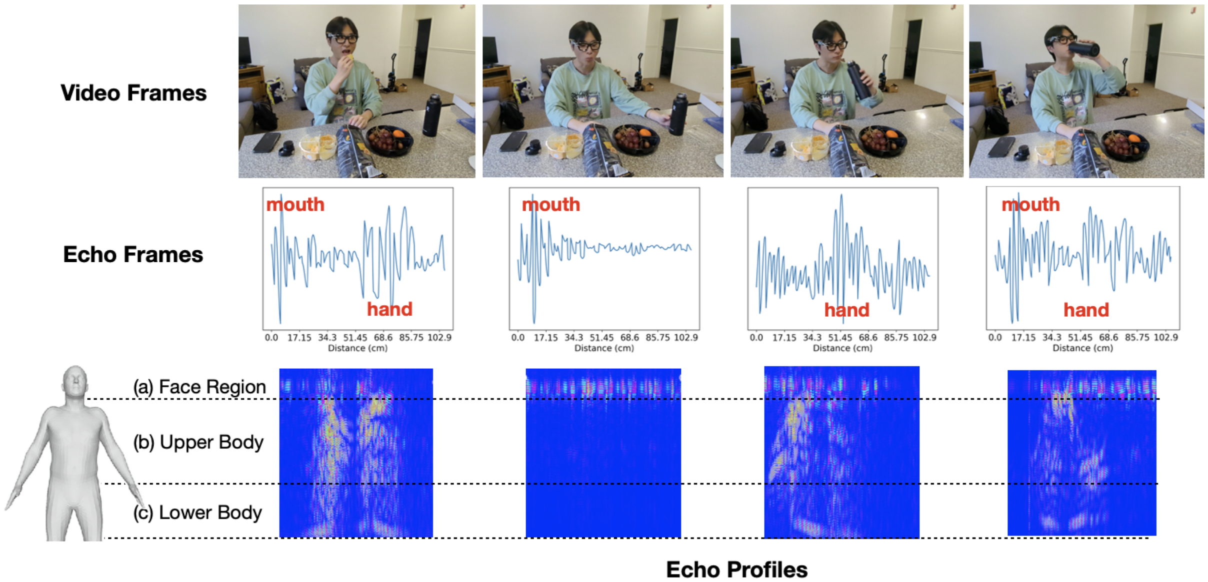

Fig. 1. Overview of the active acoustic sensing principle of ActSonic: The x-axis of the echo frames (in the 2nd row) represents the distance of echo reception. The corresponding video frames (in the first row) serve as activity references. The echo profile, created by stacking multiple echo frames, provides a spatiotemporal representation of the activity. These sliding windows with a duration of 2 seconds of echo profiles (in the 3rd row) serve as inputs for the self-supervised learning algorithm.

1 INTRODUCTION

Despite smart glasses becoming popular and being worn daily, existing eyeglasses like many other commodity wearables, still lack the ability to understand the user's fine-grained everyday activities in real-world settings. Distinguishing human activities is challenging due to the substantial variance across different activities, each involving complex and subtle movements of multiple body parts (e.g., face, arms, hands). For example, eating requires the coordination of fingers and arms to deliver food to the mouth, as well as the movement of facial muscles for chewing.

Acquiring precise body pose data from various body parts is essential for accurately distinguishing different human activities. However, achieving this with everyday wearable devices (e.g., eyeglasses) has presented persistent challenges: wearable cameras [55, 75] consume significant power, rapidly draining batteries. Furthermore, transmitting and processing video data on wearable devices is resource-intensive. Many sensing systems (Inertial Measurement Units (IMUs) [91, 109] or Electromyography (EMG) [106, 107]) only capture information around the instrumentation position, which is insufficient for activities involving multiple body parts (e.g., face and hands) unless multiple devices are worn on the body, which is less preferred by many users. Microphone-recorded passive sounds [25, 37, 66] have shown promising performance in tracking human activities in the wild. However, it does not work effectively for activities lacking distinctive auditory cues. Therefore, achieving accurate recognition of each nuanced activity often necessitates a new set of customizations on wearables with carefully selected sensor combinations, a solution that may not be generalized well across different activities. Unlike the computer vision community, which primarily relies on cameras as the sole type of sensor to record data, facilitating the establishment of various benchmark datasets, researchers in the wearable and ubiquitous computing community face challenges in collecting data to track fine-grained human behaviors.

In this paper, we aim to answer the question: *Is it possible to develop a simple sensing system on eyeglasses that is low-cost, minimally-obtrusive, energy-efficient, and privacy-aware, while capable of recognizing a wide range of everyday human activities?* A positive answer would immediately lower the barriers for researchers to study and track human activities of their interest.

To tackle this challenge, we utilized the empirical observation that many human activities involve both facial and upper-body movements. By using a single sensing system on eyeglasses to simultaneously capture high-quality facial movements and upper body poses, this information can potentially be processed by a customized deep learning framework to distinguish a wide range of human activities. However, designing such a wearable system that accurately recognizes various human activities with high resolution is highly challenging:

- The system must operate on low power to sustain continuous usage throughout the day.
- Privacy concerns of the user on wearing such a wearable system daily can be potentially alleviated with a carefully designed solution.
- The appearance of the wearable needs to be largely unchanged while wearing comfort is not compromised.
- The system must demonstrate robustness to accommodate various daily activities and distinguish them from other categories, including "Null", which refers to activities not included in the tracking set.
- Demonstrating the efficacy of the system with high temporal resolution requires recording and labeling ground truth data at the second level, a challenging and time-consuming task that many prior works overlooked.

Recent research has demonstrated the feasibility of incorporating active acoustic sensing into eyeglasses using commodity microphones and speakers. By analyzing the reflected acoustic waves from the eyes, face, or upper body, it becomes possible to infer gaze [45], facial expression [46, 110], or upper body pose [11, 28, 62] respectively. This development underscores the potential of leveraging a single sensing modality—active acoustic sensing on eyeglasses—to simultaneously capture both facial movements and upper body poses, which can be used for activity recognition. In this paper, we aim to answer the *research question*:

- Can we utilize the inaudible active acoustic signals reflected around the human body to recognize a rich set of everyday human activities?

Given the small size, relatively low power consumption, low cost, and ubiquity of microphones and speakers, such eyeglasses capable of tracking diverse human activities would significantly reduce the barriers to activity monitoring, a critical task within the ubiquitous computing community. Instead of relying on multiple devices for different activities, this all-in-one system could effectively track and differentiate between multiple activities simultaneously. However, designing such an active-acoustic sensing system and proving its efficacy for everyday activity recognition presents significant challenges:

- While active acoustic sensing has demonstrated effectiveness in recognizing facial expressions and upper body pose respectively, it remains unclear if the fused reflected acoustic signals provide adequate information to distinguish a rich set of fine-grained human activities.
- Designing machine learning and signal processing algorithms that focus on temporal and structural differences in activities from these reflected acoustic signals, rather than individual body shapes, for usability to minimize user effort for training.
- Robustness to various acoustic noises, such as multi-path acoustic reflections from the environmental setups (e.g., surrounding furniture) and potential overlap with the ultrasound band, is essential for handling different environmental conditions.
- Capturing nuanced variations in activities, particularly in large motions where finer details are significant, is crucial. For example, recognizing activities with similar arm or facial movements (e.g., yawning vs. coughing) requires an understanding of the combination of movements and temporal patterns.
- Robustness towards a large number of unknown or unlabeled activities in the "Null" category is necessary.

To address this research question, we developed ActSonic, a self-supervised and low-power activity recognition system placed on eyeglasses form factor, utilizing active acoustic sensing. It is the first system to demonstrate the feasibility of recognizing 27 types of everyday activities without collecting training data from new users. Thanks to the low-power nature of acoustic sensors, ActSonic sensing system for data acquisition can operate for over 21 hours with a battery capacity equivalent to that of Google Glass (570 mAh).

We developed ActSonic by attaching a pair of miniature, low-power, off-the-shelf microphones and speakers to the hinges of glasses. The sensing system emits inaudible ultrasonic waves to create an *acoustic aura* around the body. Based on the shape and position of various body parts, the acoustic signals are reflected with unique and fused patterns captured by the microphone. We designed a customized self-supervised deep learning framework, running on a cloud server or Android mobile phone, to interpret the reflected signals and their temporal patterns. These signals, presented with complex multipath echoes, include rich information about movements on both the face and upper body, allowing us to infer the performed activities.

ActSonic was evaluated comprehensively in two studies in real-world settings. The first study was a semi-in-the-wild investigation involving 12 participants. In this study, each participant performed all 27 activities at their own homes (12 homes) in the presence of a researcher. To further validate the system's performance in completely uncontrolled real-world conditions, we conducted a second study with 7 participants. In this second study, participants were provided with the device at their homes to record their unconstrained daily activities alone without any intervention. These two studies resulted in the collection of 40 hours of activity data from 19 different households, which were labeled with ground truth at each second. The leave-one-participant-out evaluation showed that ActSonic was able to distinguish these 27 activities at these 19 homes at each second without any training data from the user, with an average F1-score of 93.4% in the first semi-in-the-wild user study and 86.6% in the second in-the-wild study.

ActSonic advance the field of eyeglasses-based activity recognition and lower the barrier of tracking everyday human activities in the following aspects:

- The first demonstration of creating and utilizing inaudible acoustic waves around the body on eyeglasses form factor for fine-grained everyday activity recognition.
- Unlike many new data-driven activity recognition systems on wearables, ActSonic achieved promising performance without requiring training data from a new user or a new environment, which makes collecting human activity data much easier.
- Microphones and speakers are readily available, cost-effective to set up, and can sustain prolonged operation on a small battery attached to glasses, facilitating easy replication of our system by other researchers.
- Compared to many previous works relying on multiple sensors or instrumentations, ActSonic accurately recognizes 27 activities with a single sensing module at a high time resolution of one second.

2 RELATED WORK

A large and growing body of literature on activity recognition has investigated various wearable and non-wearable sensing systems, including IMUs, cameras, microphones, water pressure, and powerline sensors [13, 16], as well as multimodal sensor fusions. In this section, we provide an overview of related work focused on IMU, camera, and acoustic sensing-based activity recognition, and position the contribution of ActSonic within this landscape.

2.1 IMU-based Human Activity Recognition

Inertial Measurement Units (IMUs) in commodity smartwatches, phones, and other wearables have garnered considerable interest in detecting human activities over a significant period. These inertial sensors include accelerometers, gyroscopes, magnetometers, etc. Early research into IMU-based activity recognition relied on hand-crafted features [4, 18, 35, 76] generated from sensor readings. These methods [34] were initially limited to recognizing coarse human locomotion activities such as walking, running, sitting, etc. With the advent of end-to-end deep learning methods [15, 17, 63, 68, 73] to extract feature representations from time-domain sensor readings, IMU-based systems demonstrated an extended capability to recognize more fine-grained actions such as apparatus usage [38], body gestures [39], etc. These deep learning architectures, incorporating convolutions, recurrence, and transformers [61, 92, 100], led to the detection of activities with high fidelity and low error rates. Self-supervision for IMU-based activity recognition is particularly effective in scenarios with small labeled datasets and exhibits significant performance improvement through the utilization of large unlabeled datasets. The pre-training tasks for these self-supervised models are designed as masked window reconstruction [19], signal correspondence learning [78], and contrastive predictive coding [20] of temporal signals. Despite the success of IMU-based systems in detecting certain fine-grained activities, their capacity is limited due to the low spatial resolution [91] of the sensing modalities. Therefore, capturing a wide range of activities with a single placement of IMU remains challenging for wearables.

2.2 Vision-based and Multimodal Human Activity Recognition

Computer vision-based approaches incorporate cameras as egocentric wearables [6, 51–53, 72, 75] or systems installed in an environment [9]. In vision-based action segmentation approaches [12, 22, 33, 41, 80], they are tasked with assigning activity labels to each frame of the video. These vision-based approaches adopt weakly supervised [30] or unsupervised [33, 80] or self-supervised [75] modeling techniques to detect activities by learning temporal embedding of the video frames. On the other hand, multimodal approaches [1, 43, 65, 86, 93] utilize a fusion of sensing modalities to recognize human activities. These multimodal approaches obtain information from different combinations of IMUs, cameras, and microphones to recognize context-aware daily living activities [40, 77], body or finger gestures [60, 88, 89, 95, 103, 104]. Although vision or multimodal sensing-based activity recognition approaches demonstrate promising performance, they pose the challenge of privacy breaches and high power consumption.

2.3 Acoustic Sensing-based Human Activity Recognition

The aforementioned vision-based and multimodal sensing approaches for detecting human activities offer higher spatiotemporal resolution, leading to lower recognition errors. However, the usage of multiple sensors raises concerns about increased power consumption and privacy invasion. Recently, researchers have leveraged the pervasiveness of microphones in off-the-shelf commodity devices to recognize human activities [36, 37, 50, 96]. These acoustic sensing systems utilize passively sensed audio within the audible frequency range (20 Hz to 16 KHz). Although passively sensed environmental sound provides discriminative information required to infer certain activities that generate sound, it raises serious privacy concerns since it may record personal conversations. In response to this, existing works [5, 49, 66] adopt subsampling and other preprocessing techniques to make the audio unintelligible before activity recognition. Additionally, passively sensed audio from inaudible frequency ranges (infrasonic and ultrasonic) has been utilized to recognize daily activities [25, 26, 29]. While these techniques offer better preservation of user privacy, activity recognition on these systems relies on the assumption that activities will generate environmental sound that can be modeled by the system. However, many activities in daily living involve body-limb movement and do not necessarily generate distinctive sounds. Active acoustic sensing-based approaches emit high-frequency sound waves and leverage the reflected sounds to capture fine-grained facial movements [46, 48, 87, 108], hand poses [42, 71, 102, 104], sign language gesture [27, 28], body pose [62], activities [44], gaze [45], vital signs [70, 94], fitness monitoring [97], and robot manipulation [56, 59].

In ActSonic, we model daily living activities based on movements in different parts of the human body. To achieve this, we utilize an active acoustic sensing platform placed on eyeglasses to capture these fine-grained movements and subsequently recognize activities.

3 DESIGN AND IMPLEMENTATION OF SENSING SYSTEM

Our goal is to develop a wearable activity recognition platform integrated into glasses, enabling the identification of a wide range of everyday activities based on the positions and movements of different body parts, as captured by active acoustic sensing technology. While prior research [28, 47, 48, 62, 71, 108] has underscored the efficacy of active acoustic sensing in detecting facial changes and upper body poses, no existing system has seamlessly integrated this sensing modality for the comprehensive task of classifying diverse daily activities in real-world scenarios. Given that most daily activities entail movements across various body parts, particularly the upper body, head, and face, our primary focus with ActSonic is to capture and differentiate these activities based on the position and movement of these body areas, as detected by our sensing system. ActSonic accomplishes this with a singular sensor placement, facilitating concurrent information capture. This section provides an overview of the active acoustic sensing setup, the process of feature extraction for activity recognition, and the hardware implementation, with a particular emphasis on the wearable form factor of the system.

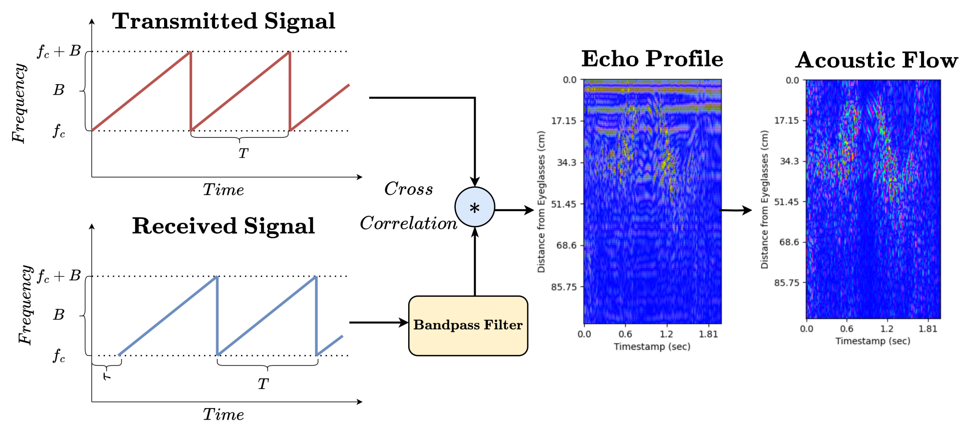

Fig. 2. Overview of echo profile and acoustic flow calculation. For the echo profile, we cross-correlate the transmitted signal with a bandpass filter applied over the received signal (to ensure only specific frequencies are returned). This allows us to capture the direct echo profile, and we can calculate acoustic flow by taking the difference between two consecutive echo profiles.

3.1 Configuration of Active Acoustic Signal

The active acoustic sensing system in ActSonic incorporates two pairs of ultrasonic transmitters and receivers on the eyeglass hinges. Utilizing Cross-correlation-based Frequency Modulated Continuous Wave (C-FMCW) [94] chirps, ActSonic emits ultrasonic signals with linearly modulated frequencies ranging from 18.0 to 21.0 kHz and 21.5 to 24.5 kHz simultaneously for the left and right transmitters, respectively, each with a bandwidth of $B = 3$ kHz.

For the C-FMCW [94] technique employed by ActSonic, the theoretical range resolution R can be derived as follows:

$$R = \frac{C \cdot Lag}{2F_s} - vt \tag{1}$$

where C is the speed of sound ($C = 343$ m/s in dry air at 20°C), F_s is the sampling frequency of the microphone or receiver ($F_s = 50$ kHz in the case of ActSonic), v is the velocity of the tracked body part, t is the elapsed time from the start of the modulation period, and Lag is the number of samples shifted between the transmitted and received acoustic signals. With $N = 600$ samples transmitted for each chirp, ActSonic's sensing system has a sweep period of $T = \frac{600}{50000} = 0.012$ seconds = 12 milliseconds. Thus, the value of t ranges from $0 \le t \le 0.012$ s, and the value of Lag in Eq. 1 ranges from $0 \le Lag \le 600$.

According to the rationale provided in [94], we can set $v = 0$ in Eq. 1 in the case where the tracked body part is static or moving slowly with respect to the speed of sound. The velocity of extremely fast arm movements, such as punching, has been measured at approximately 8.9 to 11.5 m/s [31], while the duration of the fastest facial expression in humans is measured at 100 ms [99]. The speed of these extreme body motions (v) is significantly smaller compared to the speed of sound (C). Therefore, based on the equation above, the impact of the sensing resolution (vt) caused by body movements is minimal. As a result, the theoretical range resolution δR of the C-FMCW signal deployed in ActSonic can be simplified to:

$$\delta R = \frac{C \cdot \delta Lag}{2F_s} \tag{2}$$

Since cross-correlation is computed for each sample in the case of C-FMCW [94], we can set $\delta Lag = 1$ for the theoretical range resolution calculation in Eq. 2. Therefore, we can substitute the values of variables in Eq. 2 and compute the resolution of the sensing system of ActSonic as follows:

$$\delta R = \frac{343 \text{ m/s}}{2 \cdot 50000} = 0.343 \text{ cm} = 3.43 \text{ mm} \tag{3}$$

The maximum sensing range of ActSonic can be computed as $R_{\max} = N \cdot \delta R = 600 \cdot 0.343 \text{ cm} = 205.8 \text{ cm} \approx 2$ m. This combination of high sensing resolution (0.343 cm) and extensive sensing range enables us to capture both subtle skin deformations on the face and monitor the pose and coarse movement of the upper body region effectively. Additionally, it is important to note that the sensing resolution mentioned above is calculated for range-finding applications, as demonstrated in [94], which differs from our approach. This calculation is provided as a reference to facilitate readers' understanding of the sensing principle underlying our proposed system.

3.2 Computation of Echo Profile and Acoustic Flow

Active acoustic sensing mechanism in ActSonic measures the round-trip delay between emitted and reflected ultrasonic waves to detect human body movements. To capture this delay (τ), we utilize cross-correlation [94] on transmitted and received signals. Figure 2 demonstrates the application of bandpass filters matching the transmitted frequency ranges (18 − 21 KHz and 21.5 − 24.5 KHz) on received signals. This filtering eliminates audible frequencies, ensuring user privacy and eliminating environmental acoustic noise before cross-correlation computation.

The cross-correlation matches the sweep period of the emitted FMCW ultrasonic wave. It generates an *echo frame*, represented as a (600 × 1) column vector in ActSonic. Stacking these frames creates the *echo profile* [48, 71, 94, 108], where brighter pixels indicate strong reflections at specific distances. With two transmitter-receiver pairs, ActSonic accounts for four transmission paths. To elaborate, if we consider the left and right transmitter-receiver pair as (T_{left}, R_{left}) and (T_{right}, R_{right}) respectively, then the paths will be $T_{left} \rightarrow R_{left}$, $T_{left} \rightarrow R_{right}$, $T_{right} \rightarrow R_{left}$, and $T_{right} \rightarrow R_{right}$. In ActSonic, we stack the outputs of cross-correlation of these four paths as four channels of the echo profile. Note that the computation of one channel in the echo profile is shown in Figure 2.

Acoustic flow, also known as the differential echo profile, is derived by computing the derivative of distance from the eyeglasses (echo profile's y-axis) with respect to time (x-axis). This is achieved by calculating the absolute difference between consecutive echo frames. Acoustic flow effectively eliminates reflections from stationary objects, enabling precise detection of human body movements. Moreover, it mitigates the effects of eyeglasses remounting, ensuring a resilient measurement of body motion across sessions. The y-axis of the echo profile (from top to bottom) indicates the distance from the eyeglasses, while the x-axis represents the temporal axis. The sliding window employed in ActSonic covers a sensing range of 300 pixels, roughly equivalent to 1 meter, extending up to the user's knees. This parameter of 300 pixels has been fine-tuned to optimize activity recognition.

3.3 Hardware Implementation and Wearable Form Factor

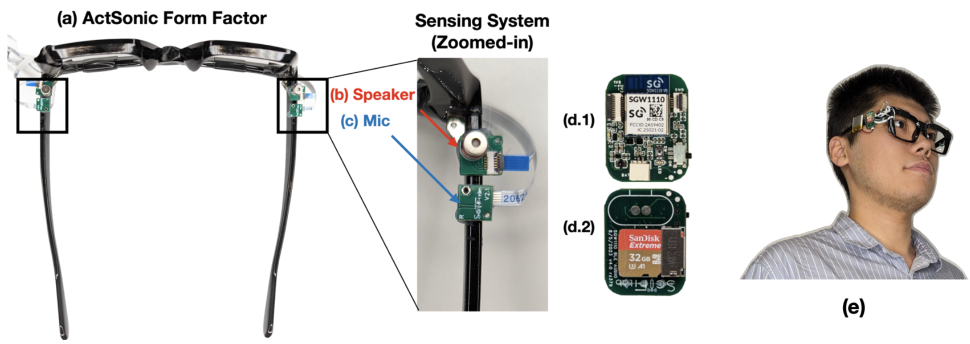

Fig. 3. Hardware of ActSonic: **(a)** Eyeglasses form factor, **(b)** Transmitter or speaker, **(c)** Receiver or microphone (dimension of the sensor board of (b) and (c) is $9mm \times 9mm$), (d) Front (d.1) and back (d.2) of customized PCB board (dimension $18mm \times 23mm$) with low-power nRF52840 micro-controller, (e) User wearing ActSonic eyeglasses form factor

We assembled the active acoustic sensing system for ActSonic using two OWR-05049T-38D speakers and two ICS-43434 microphones [90], following a design similar to that shown in [48]. Managed by a Teensy 4.1 microcontroller [74], the setup oversaw FMCW signal transmission and reception. To connect the speakers, microphones, and microcontroller, we developed a custom PCB housing two MAX98357A audio amplifier chips [64]. Utilizing the Inter-IC Sound (I2S) interface, ActSonic's hardware components communicate, with received signals stored on an SD card via the micro-SD interface on the microcontroller.

Positioned symmetrically on a standard pair of glasses, the ActSonic sensor system optimally captures nuanced body movements from various angles, facing perpendicularly downwards towards the body. After several design iterations, this orientation proved most effective for sound wave propagation. Connected via Flexible Printed Circuit (FPC) cables, the microcontroller, along with a Li-Po battery, is affixed to one leg of the glasses, interfacing with the speakers and microphones.

Initially, our prototype utilized the Teensy 4.1 [74] microcontroller, powered by an ARM M7 core, which exhibited higher power consumption due to its characteristics. This Teensy 4.1 [74]-based prototype was used to collect data in the user studies intended to evaluate the system. To highlight the power efficiency of our acoustic sensing, we designed a low-power variant featuring an nRF52840 microcontroller [82], based on a low-power ARM M4 core. This variant includes two MAX98357A audio amplifiers, similar to the original setup, alongside power management modules and the SGW1110 [83] module. A 32 GB SanDisk Extreme microSD card [3] handles storage, with optimized firmware minimizing SD card accesses for swift operations.

4 DEEP LEARNING FRAMEWORK

To estimate human activities from acoustic sensing data in ActSonic, we design a self-supervised deep learning framework which is illustrated in Figure 4. This framework leverages the unlabelled acoustic data to create a pre-trained encoder. Then, we use this encoder to create an activity recognition pipeline.

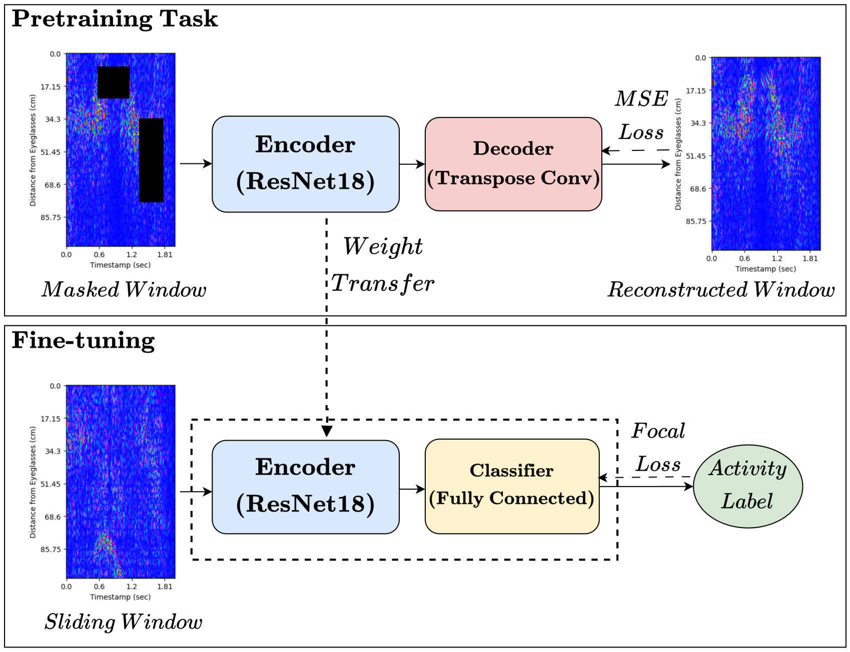

Fig. 4. Deep learning model architecture for ActSonic. Within the self-supervised **pretraining** stage, we mask out specific sections of the input echo profile and train an encoder-decoder architecture to reconstruct the input echo profile (given a lightweight decoder) supervised by an MSE loss. We then **fine-tune** the trained encoder from this step along with a lightweight classifier on the labeled dataset.

4.1 Self-supervised Learning Pipeline

Self-supervised learning is a form of supervised learning where the model predicts a subset of unlabelled data from the rest. This learning pipeline of ActSonic consists of two steps: **pretraining** encoder to learn the representation of unlabelled data, and **fine-tuning** the pre-trained encoder weights for the target task with labeled data.

4.1.1 Pretraining Task. The pre-training approach to learning representation from the unlabelled data is to perturb the sliding window of acoustic flow (described in 3.2) with binary mask and reconstruct the original window using the autoencoder depicted in the pre-training task segment of Figure 4. Here, the binary mask is constructed in a way such that $m\%$ of each channel of the sliding window is set to 0. In the case of ActSonic, the numerical value m of this mask percentage is randomly chosen from the range $15 - 20\%$, and the number of patches is randomly chosen from the range 1 to 4. As illustrated in Figure 4, the masked window goes through a ResNet18 [21] encoder. The embedded representation from the encoder is then fed through a decoder network, essentially a transpose convolution network. This customized transpose convolution, or deconvolution, network takes feature maps from the ResNet18 as input and generates a three-dimensional matrix of the shape of the input sliding window. We calculate the Mean Squared Error (MSE) between each pixel in the reconstructed and original sliding window as the loss function for this autoencoder architecture.

4.1.2 Fine-tuning. The aforementioned ResNet18 encoder of the pre-training pipeline (detailed in Subsec. 4.1.1) learns the representation of active acoustic data via self-supervision. This ResNet18 encoder with learned weights serves as the feature extraction pipeline in the fine-tuning phase. We design an activity recognition architecture (depicted in the fine-tuning segment of Figure 4) comprising of pre-trained ResNet18 encoder followed by a fully connected classifier layer. We apply average pooling on the spatial axis of the feature map extracted by the encoder and feed it to the fully connected layer. The fully connected classifier network is a feedforward neural network with batch normalization [24], Leaky ReLU activation [98], and dropout [85] in between. We set the number of neurons in the last layer equal to the number of activity classes and perform a softmax operation to output a probability distribution.

The activity recognition model in the fine-tuning phase is trained using acoustic flow sliding windows as input and activity class labels as the target. To optimize the training process, we employ focal loss [54] as the objective function. Focal loss is a modification of the standard cross-entropy loss, designed to emphasize learning from hard examples. It dynamically scales the loss function based on the confidence of the correct class prediction, with a decay factor that decreases as the confidence increases. In binary classification scenarios, where p_t represents the predicted class probability, the standard cross-entropy loss $CE(p_t)$ is defined as $-\log(p_t)$ when $p_t = p$ for the positive class and $-\log(1 - p_t)$ otherwise. Focal loss adds a scaling factor $(1 - p_t)^\gamma$ to this standard cross-entropy loss, with γ being a hyperparameter set to 0.5 for ActSonic. This modification ensures that the loss function assigns lower values to well-classified examples ($p_t > 0.5$) and focuses more on misclassified examples. This adaptation is particularly effective for ActSonic due to the imbalanced distribution of activity labels in the dataset and the similarity in body motion patterns observed in the echo profile for some activities.

4.2 Training and Implementation

The self-supervised activity recognition model of ActSonic processes overlapping sliding windows of the acoustic flow as input, with the shape of the sliding window being a hyperparameter. We conduct an iterative process to determine the optimal sliding window duration, ranging from 0.30 seconds to 5.00 seconds with a hop size of 0.10 seconds. Performance evaluation on the validation set helps us fine-tune this parameter, resulting in an optimal duration of 2.00 seconds with a 50% overlap. The shape of the input sliding window is defined as (num_channels × num_features × num_samples) = (4 × 295 × 166). Here, num_features represents the number of pixels from the echo profile, calibrated to 295, covering a sensing range of approximately 1 meter (precisely 101.185 cm), sufficient to capture upper body poses. Considering the sampling rate of ActSonic at 50 KHz and the number of samples in one sweep period at 600 (details in Sec. 3), a one-second sliding window contains approximately $\lfloor \frac{50000}{600} \rfloor = 83$ samples. Consequently, the num_samples for a 2.00-second sliding window of ActSonic is set to 166.

The dropout probability of the feedforward classification layer in the fine-tuning phase is configured to 0.2. Both the pre-trained and fine-tuning models are trained for 100 and 50 epochs, respectively, using a batch size of 64. We employ the Adam optimizer [32] and incorporate a cosine annealing learning rate scheduler with an initial learning rate of 10^{-3}. The self-supervised model, including both the pre-training and fine-tuning networks, is implemented using the PyTorch and PyTorch Lightning frameworks and trained on GeForce RTX 2080 Ti GPUs.

4.3 Evaluation Metric

We use the Macro F1-score as our evaluation metric for ActSonic's activity recognition performance. If C is the set of activity classes such that classes are indexed as $0, \ldots, (C-1)$ and $|C|$ is the cardinality of this set, the evaluation metric is defined as:

$$Macro\ F1 = \frac{1}{|C|} \cdot \sum_{i=0}^{C-1} \frac{2 \cdot precision_i \cdot recall_i}{precision_i + recall_i} \tag{4}$$

Where $precision_i$ and $recall_i$ are the numerical values of precision or positive predictive value and recall or sensitivity of i-th class respectively.

5 USER STUDY

In this section, we present a comprehensive overview of the user studies conducted to assess the performance of ActSonic. The objective of these studies is to evaluate the activity recognition pipeline under naturalistic conditions. To achieve this goal, we devised a diverse set of everyday activities to monitor throughout the study, recruited participants, and carried out both a semi-in-the-wild user study and a more extended unconstrained study, both conducted in participants' homes in a naturalistic environment.

5.1 Design of Activity Set

To establish a set of activities, we conducted a pilot feasibility study encompassing over 50 activities of daily living with 5 users from our research team. Drawing on relevant prior studies [10, 25, 37, 66] and insights from the pilot study, we selected 27 activities of daily living to be incorporated into ActSonic's tracking set. Additionally, we introduced a *Null* label for activities not part of the tracking set. The primary criterion for selecting everyday activities was their involvement of movements across different body parts, aligning with ActSonic's reliance on tracking body motion. The activities in the tracking set are categorized into three segments based on their typical indoor locations:

- **Bathroom (6 activities):** Rinse Mouth, Brushing Teeth, Flossing, Brushing Hair, Flush Toilet, Opening Door
- **Kitchen and Dining Area (8 activities) :** Washing Hands, Eating, Drinking, Pickup / Putdown, Pouring, Chopping, Wiping Surface, Stirring
- **Bedroom and Living Area (13 activities):** Stationary, Walking, Sitting, Coughing, Yawning, Talking, Putting on Outerwear, Vacuum, Throwing, Stretching, Using Phone/Tab, Squat, Reading Book

5.2 Participants and Study Schedule

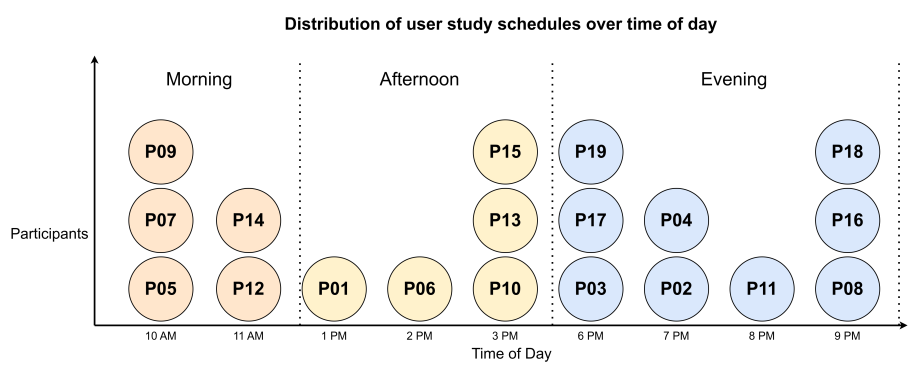

Fig. 5. Distribution of participant schedules for the user study over time, where x-axis represents time and y-axis represents participant count. We split the participants into three general groups ("morning" as 7 am - 12 pm, "afternoon" as 12 pm - 6 pm, and "evening" as 6 pm - 11 pm) and ensure that we get a mixture of data across different times of day

The ActSonic user studies received approval from the Institutional Review Board for Human Participant Research (IRB) at our organization. We enlisted 12 participants for a semi-in-the-wild study and 7 different participants for an unconstrained user study. Among the total 19 participants, with an average age of (24.737 ± 3.445) years, ranging from 21 to 31 years, 5 identified as female, and 14 identified as male. Per IRB guidelines, each study lasted no longer than 2 hours (120 minutes), and participants received $30 USD as compensation for their time. Post-study, we gathered basic demographic and physical information (e.g., height, weight, gender), along with general feedback on the ActSonic wearable device via an IRB-approved questionnaire.

The user studies took place in participants' homes, where they utilized their own tools or appliances as needed for activities. An exception was made for a few participants who were provided with dental floss for the flossing activity. A trained experimenter from our research team visited participants' addresses equipped with the necessary data collection apparatus to conduct the study.

Based on insights from a pilot feasibility study, human activity patterns vary throughout the day. For instance, activities such as tooth brushing are more likely to occur in the morning or after dinner. Accordingly, we scheduled user studies to capture activity data across different parts of participants' daily routines. Users provided their preferred study date and time slots for each part of the day (morning, afternoon, and evening as defined in Figure 5). As shown in Figure 5, each point represents the start time of a study session, and the respective users were randomly assigned based on their preferences.

5.3 Data Capture Apparatus

We captured acoustic data using the sensing system integrated into ActSonic eyeglasses. Additionally, we recorded ground truth video data to annotate the activities. For this purpose, we employed a GoPro HERO9 camera [14] mounted on the participants' chests using a lightweight body mount from the same manufacturer. The camera's horizontal and vertical field of view was set to 118° and 69° respectively. It recorded egocentric videos at a resolution of 720p and a frame rate of 30 fps. Additionally, participants were provided with Apple AirPods Pro during sessions where they received audio prompts or instructions for specific activities.

5.4 Study Design

We conducted a 12-participant *semi-in-the-wild study* followed by an *unconstrained study* with 7 participants. Both of these studies were conducted at participants' homes in unconstrained settings. The design of the study protocols is discussed below.

5.4.1 User Study Design Considerations. We conducted two separate studies at participants' homes in a naturalistic setting. Given that the activity set of ActSonic encompasses a wide range of activities, obtaining samples of each activity in an uncontrolled setting, where participants are allowed to carry out their daily routines, proved challenging. Therefore, we conducted a semi-in-the-wild study where users were prompted with specific instructions to perform activities, thus covering the tracking set of ActSonic. Subsequently, we conducted a second unconstrained study at participants' homes, where they did not receive any particular instructions to perform activities but instead continued with their daily routines.

We utilized a chest-mounted camera to gather reference videos for ground truth annotation. In contrast to other activity recognition systems evaluated in real-world settings [84], we opted not to rely on self-reporting-based annotation, despite its facilitation of longer duration in-the-wild data collection without significant labeling effort. This decision stems from the impracticality of accurately self-reporting all 27 types of activities throughout a day. Additionally, we aimed to evaluate ActSonic based on the inference generated at each second, necessitating fine-grained ground truth annotation from the uncontrolled study, which is only achievable with a body-mounted camera.

5.4.2 Study - 01: Semi-in-the-wild User Study. The semi-in-the-wild user study is partitioned into two segments. In the first segment, the participants received audio instructions to perform certain activities by wearing Apple AirPods Pro. The goal of this study segment is to collect data samples of all the activities included in the recognition set of ActSonic. Before starting this segment of the study, the participants were briefed about the procedure and familiarized with the audio instructions they were going to receive for each activity.

The activity set was split according to the indoor locations mentioned in Subsec. 5.1. Two sessions of activity data were collected for the bathroom and kitchen locations. The living area activities were divided into two subsets for the convenience of participants, and two sessions of data were collected for each subset. In each session of the data collection process of this segment, the participants received audio instructions for specific activities in random order, and each activity was repeated 5 times within the session. The duration of each repetition of activities spanned from 10 to 30 seconds. The participants were provided minimal instruction regarding the way to perform certain activities so that they could perform their natural body movements. The duration of this segment, comprising a total of eight sessions, is 68 minutes, and each session was 8.5 minutes long. In between each session, the participants were asked to remount the ActSonic eyeglasses.

In the second segment of the semi-in-the-wild study, the participants did not receive any prompt or instruction to perform certain activities. They were allowed to perform their regular daily routine. The total duration of this segment was 30 minutes and was divided into three sessions. The participants wore a chest-mounted camera so that the ground truth video could be recorded for activity annotation. When the participants were performing the activities in this segment, the experimenter was present at the participants' home.

5.4.3 Study - 02: Unconstrained Study at Participants' Home. We designed a longer-duration unconstrained study with the goal of evaluating the ActSonic activity recognition system in the wild for extended hours. Furthermore, the activity set of ActSonic (listed in Subsec. 5.1) contains a total of 27 activities, including the *null* class, and it is unlikely to get samples of all those activities in the second segment (30 minutes) of the semi-in-the-wild study in a naturalistic setting. Since the IRB protocol allows a maximum duration of two hours for a single study, we designed this in-the-wild study with a duration of two hours. However, the protocol allows multiple studies with the same participant, and therefore the participants were given an option to take part in multiple consecutive studies. One out of the 7 participants opted for that choice and participated in two consecutive studies which were four hours long in total, and they were compensated twice. Hence, we accumulated 16 hours of in-the-wild data from this study.

We followed the same data collection procedure for the unconstrained study as the second segment of the semi-in-the-wild study. One difference to be noted is that the experimenter left the participants' homes after briefing them and setting up the data collection system. The participants returned the data collection system after two hours. In this unconstrained study, the participants were not instructed with any specifics of the activities to be performed during the study; rather, they were allowed to continue their regular schedule at their homes. Note that we did not permit participants to leave their homes due to legal restrictions on video recording in public places in the country where the study was conducted.

5.5 Peer-reviewed Data Annotation Protocol

To provide annotations for active acoustic data via ground-truth egocentric video data, we utilized the ANU-CVML Video Annotation Tool (Vidat) [105] to annotate all ground-truth egocentric video data with action annotations. Subsequently, we synchronized the timestamps of the acoustic and video data using a clapping action at the beginning of the recording, which had a distinct acoustic signature. We then developed separate postprocessing scripts to align video annotations with acoustic echo profiles. To ensure accurate annotation of activities in a naturalistic setting, we implemented a peer-reviewed annotation process. In this procedure, one annotator from the research team labeled the data, while another researcher independently reviewed the annotations and provided feedback. After a phase of revision and approval from the reviewer, the annotated labels were incorporated into the ActSonic dataset.

6 PERFORMANCE EVALUATION

We evaluated the performance of ActSonic on the data collected in user studies. Our evaluation can be partitioned into two phases. In the first phase, we evaluate the activity recognition performance on the prompted sessions of the semi-in-the-wild user study. Subsequently, we benchmark the performance of ActSonic on unconstrained sessions of the user study. For both scenarios, we employed a leave-one-participant-out strategy, to evaluate its performance without the need to collect training data from any new user in new environments.

6.1 Leave-one-participant-out Evaluation of Prompted Sessions

We conducted a leave-one-participant-out cross-validation evaluation on prompted sessions of the semi-in-the-wild study. In the initial segment of the semi-in-the-wild study (as described in Sec 5), involving 12 participants, each participant contributed 8 sessions. We trained 12 user-independent models, where, for instance, the model for participant P01 was trained solely on data from P02-P12 and tested on P01, and the same process was repeated for the other participants. The average macro F1-score for each participant ranged from 0.90 to 0.95, exhibiting a standard deviation of 0.035. Across all participants, the average macro F1-score in the leave-one-participant-out evaluation was 0.934. When examining individual activities, we observed high accuracy (macro F1-score) for all activities, ranging from 0.88 to 0.97 across participants in the semi-in-the-wild study with a prediction frequency of 1 Hz. Compared to prior work[36, 66], which evaluated activity recognition in a similar manner (albeit with different activities), ActSonic has significantly better performance with a wider range of activities. However, we also acknowledge that in this part of the study, participants performed activities following audio stimuli with the presence of a researcher at their home, which may reduce the variance in how they perform activities in real-world settings.

6.2 Evaluation of User Independent Model on Unconstrained Sessions

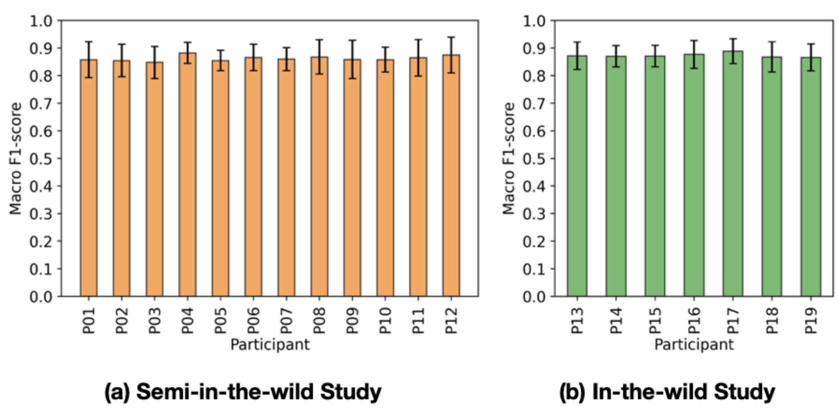

Fig. 6. Leave-one-participant-out performance evaluation of **unconstrained sessions**

We further assess our user-independent models through a leave-one-participant-out cross-validation strategy on our dataset of unconstrained sessions. This evaluation aims to gauge ActSonic's performance in real-world naturalistic scenarios. As noted in Sec. 5, participants continued their regular daily routines at home during the study. Hence, these sessions entail activity samples that might exhibit more diverse motion profiles compared to prompted sessions.

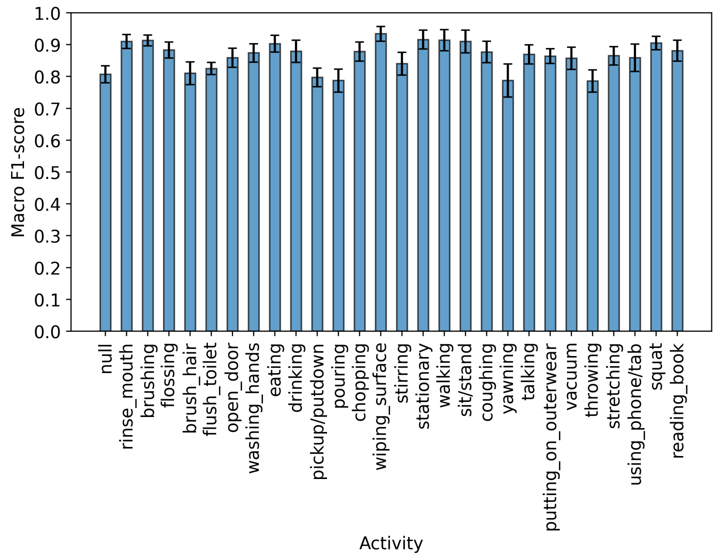

Fig. 7. Evaluation of the performance of different activities in **unconstrained sessions**

Our evaluation of the unconstrained sessions from both the semi-in-the-wild and the second studies involves two stages. In the initial stage, we take each model trained on prompted sessions and conduct a leave-one-participant-out evaluation (as described in Subsec. 6.1). Subsequently, we fine-tune these models using unconstrained data from the semi-in-the-wild study for other participants before assessing the model on the original participant. For instance, the P01 "prompted" model, trained on P02-P12 "prompted" supervision, undergoes fine-tuning using "unconstrained" supervision from P02-P12, ensuring the exclusion of labels from P01 during model training.

In the second phase of the unconstrained session evaluation, we measure the performance using data from the second study. To assess the unconstrained sessions of users P13 to P19 from the unconstrained study, we first train a fine-tuned model using prompted session data of users P01 to P12 as supervision. Subsequently, we evaluate the model performance using data from individual users of P13-P19 as the test set.

The average macro F1-score and standard deviation for all actions listed in Sec 5.1 are reported in Figure 6. Specifically, Figure 6(a) and 6(b) present the average activity recognition performance on semi-in-the-wild and unconstrained study participants respectively. Furthermore, Figure 7 displays the average macro F1-score and standard deviation for each of the unconstrained actions across all participants. Our findings reveal an average F1-score of 0.866 with a standard deviation of 0.052 with a predication frequency of 1 Hz. Although the system's performance in an unconstrained scenario is comparatively lower than in prompted sessions, it is apparent in Figure 9 that the activity recognition remains accurate, considering the user variability and class imbalance present in the unconstrained sessions.

The class distribution, indicating the number of participants performing specific activities, is illustrated in Figure 8. The observed accuracy imbalance among different actions within the unconstrained examples likely stems from the natural variability in individual interactions within an open-world setting, influenced by external context. While activities like "rinsing mouth," "brushing," "walking," and "sitting/standing" seem less contextually dependent, actions such as "pouring," "throwing," and "pickup/putdown" might display lower performance due to subtler movements and inherent variability arising from interactions with different objects.

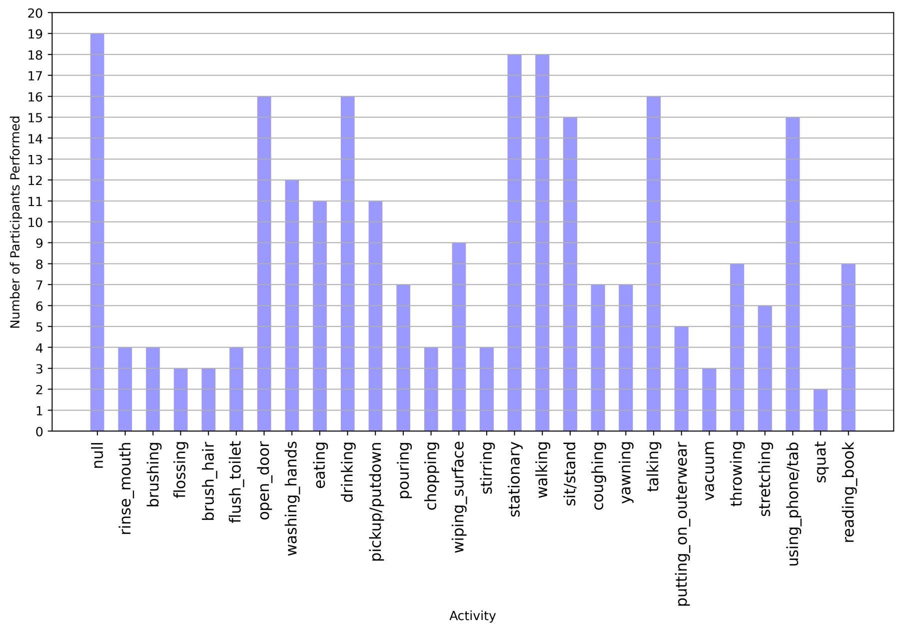

Fig. 8. Number of participants per activity during **unconstrained sessions**: activity labels on the x-axis and number of participants on the y-axis

6.3 Power Signature of the ActSonic Sensing System on Eyeglasses Form Factor

ActSonic system initially consumed 577.8 mW for collecting data with the first prototype employing Teensy 4.1. Substituting this microcontroller with a low-power nRF52840 significantly reduced power consumption. We integrated the original speakers and microphones from the initial prototype into the second one, adjusting the gain to ensure identical sound pressure levels (SPL) for both, maintaining emission consistency. The power signature, measured using a CurrentRanger [58], displayed an average operation of 96.5 mW (4.02 V, 24.0 mA) while saving all data to the SD card. Furthermore, we conducted long-term stability testing, and the prototype operated continuously for 11.3 hours using a 290 mAh 3.7 V Li-Po battery. This configuration enables a full-day operation on commodity smart glasses or AR glasses. For instance, Google Glass, equipped with a 570 mAh battery, can support the ActSonic sensing system for over 21 hours if the activity recognition pipeline is the only active process.

6.4 Latency and System Overhead of ActSonic Inference Pipeline on Mobile Platform

We evaluated the latency and overhead of the ActSonic inference pipeline on the Google Pixel 7 mobile platform. Table 1 presents the inference time and various parameters of the ActSonic ResNet18 model evaluated on an unconstrained dataset with the same protocol described in 6.2. Initially, we generated lightweight mobile models from both the original and the 8-bit integer quantized versions of the ResNet18 model using PyTorch Mobile. Subsequently, we conducted inference time benchmarks using a two-second sliding window on a Google Pixel 7 Android phone, performing inference on 1000 samples. The mean inference time is detailed in Table 1. The ActSonic performance on mobile devices was evaluated by transmitting the acoustic data to a Pixel 7 for inference. The BLE status was set to "connected" since it continuously streams data to mobile devices.

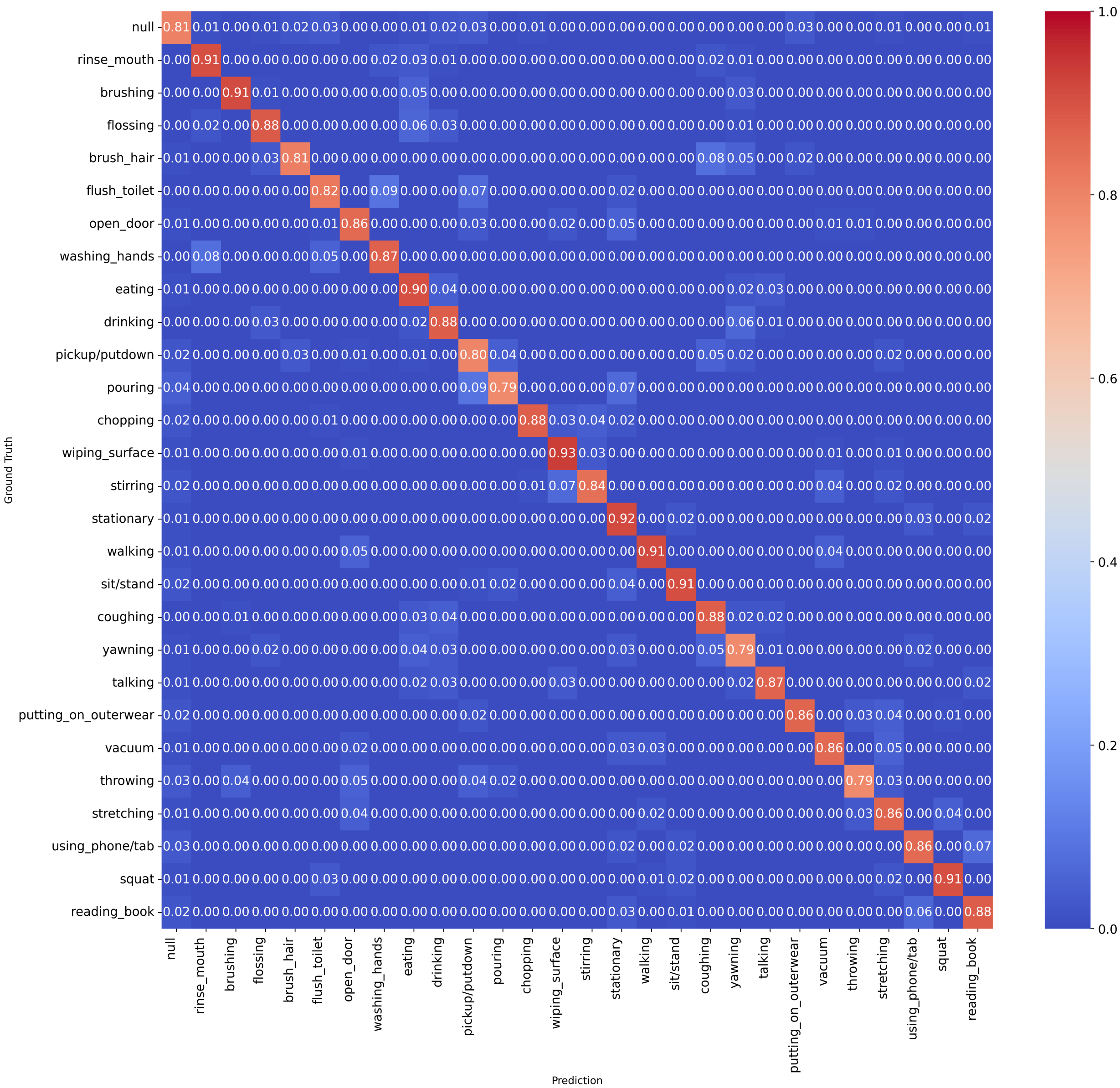

Fig. 9. Normalized confusion matrix of leave-one-participant-out user evaluation in **unconstrained sessions**

Furthermore, we utilized the Android Profiler [8] to evaluate mobile CPU usage and energy consumption. Table 1 indicates that while the quantized model exhibits lower accuracy compared to its non-quantized counterpart, it demonstrates lower system overhead. This assessment utilized a post-training quantization strategy; employing a quantization-aware training approach for the ResNet18 model might yield improved performance while preserving

	Non-quantized	Quantized
Avg. Macro F1-Score	0.864	0.772
Size	151.1 MB	45.4 MB
Inference Time	123.1 ms	68.4 ms
CPU	14%	11%

Table 1. ActSonic ResNet18 model latency and system overhead on Google Pixel 7 android mobile platform.

similar efficiency. The energy consumption was indicated as "light" for both the quantized and non-quantized models by the Android Profiler [8].

7 ABLATION STUDY

7.1 Impact of Sensing Different Body Parts on Activity Recognition Performance

ActSonic relies on tracking the movement of facial and upper body limbs to recognize everyday activities. To assess the impact of different body regions on activity recognition performance, we conduct an evaluation of the ActSonic system using acoustic signals solely from the face and upper body regions. We then compare this recognition performance with the evaluation reported in Sec. 6. In order to filter out the acoustic reflection from the face region, we crop the first 50 pixels from the top to bottom of the y-axis of the echo profile sliding window. This 50 pixel (= 17.15 cm) approximately represents the face region of the user and the movement from this region is captured in the cropped differential echo profile. In addition, we evaluate the performance of ActSonic with the rest of the echo profile sliding window (representing the upper body region up to the knees). We present the performance of ActSonic under these scenarios in Figure 10.

From the performance reported in Figure 10, we observe a sharp degradation in performance if we exclude the movement patterns of the upper body region. Analyzing the activity-wise performance, we note that activities involving obvious facial movements have fewer errors compared to activities that involve upper body movements, such as eating, drinking, talking, etc. On the other hand, we observe less degradation in performance if we exclude the face region movement from the acoustic signal. This observation can be attributed to the fact that most activities in the ActSonic dataset involve hand or upper body movement. Overall, based on the evaluation presented in Figure 10, we can extrapolate that the combination of reflection patterns from the face and upper body regions yields the best performance for the ActSonic system.

7.2 Impact of Different Placements of the ActSonic Sensing System on the Eyeglasses Form Factor

We evaluated the impact of sensing system placement on the performance of ActSonic in this experiment. The raw audio data was collected in user studies using the ActSonic sensing system placed on both hinges of the eyeglasses. To measure the impact of sensor placement, we utilized the acoustic data from one side, either left or right. We then computed the echo profile and acoustic flow using that data. With only one side activated, we have just one C-FMCW frequency range, yielding echo profiles with one channel containing information on the direct acoustic transmission path (details of transmission paths are mentioned in Sec. 3.2). We customized the deep learning pipeline to take input sliding windows of the same shape with one channel and evaluated the activity recognition performance of ActSonic.

Figure 11 illustrates that placing the sensing system on both hinges of the eyeglasses yields the best performance, with a mean macro F1 score of 0.866 ± 0.044 in our leave-one-participant-out evaluation. The right-side placement demonstrated moderate performance, with a mean macro F1 score of 0.838 ± 0.057, which is close to the performance of both-side placements. This can be explained by the involvement of the dominant hand in most of

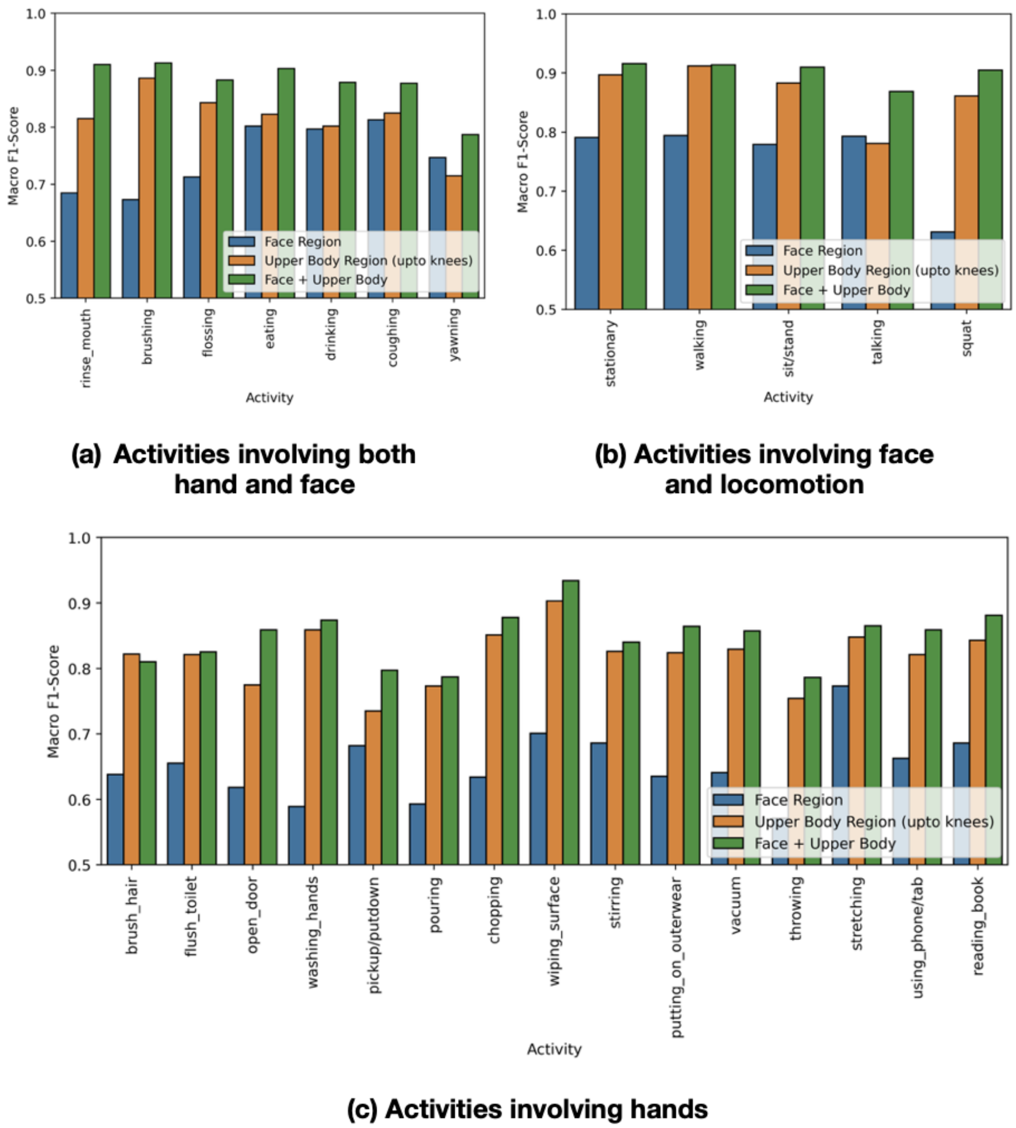

Fig. 10. The impact of using acoustic signals corresponding to different body regions on the performance of ActSonic is evaluated. The face region comprises the first 50 pixels in the differential echo profile, covering movement within 17.15 cm of the sensing system. The upper body region encompasses the remainder of the echo profile sliding window. "Face + Upper Body" denotes the performance of ActSonic using the entire sliding window (whose shape is tuned as a hyperparameter).

the activities in the ActSonic recognition set, which was the right hand for all the participants in the study. On the other hand, the left hinge placement demonstrated degraded performance compared to the right one, with a mean macro F1 score of 0.811 ± 0.062.

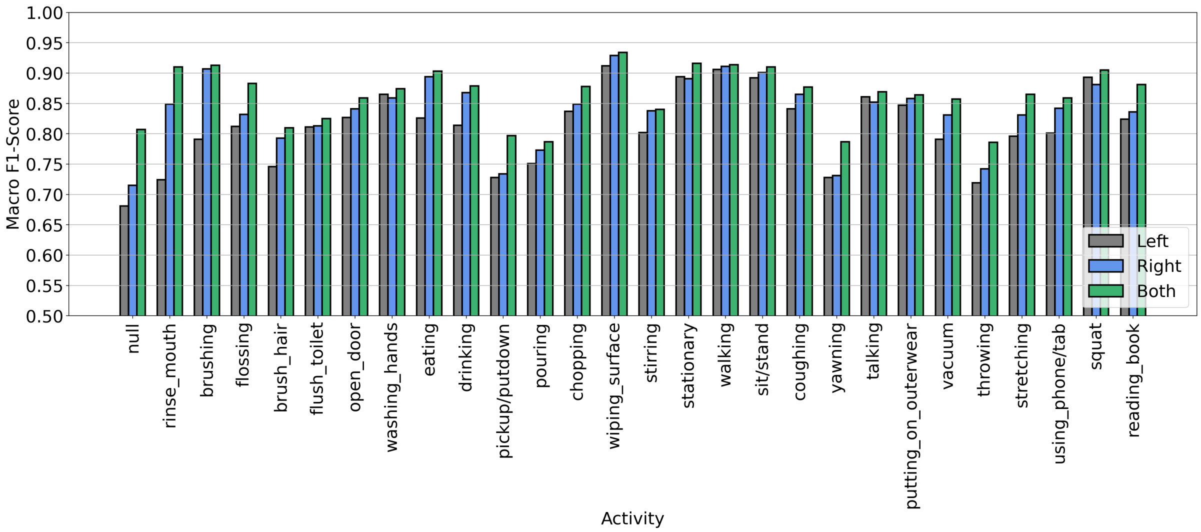

Fig. 11. Impact of active acoustic sensing placement on recognition performance of ActSonic: **Left** indicates a speaker (emitting C-FMCW chirps within the frequency range of 18.0-21.0 KHz) and microphone pair on the left hinge of the eyeglasses, **Right** indicates a speaker (emitting C-FMCW chirps within the frequency range of 21.5-24.5 KHz) and microphone pair on the right hinge of the eyeglasses, and **Both** indicates both aforementioned speaker-mic pairs are placed on both hinges of the eyeglasses.

7.3 Performance Comparison of Different Deep Learning Encoders

Model Architecture	Number of Parameters (Approx.)	Prompted Sessions	Unconstrained Sessions
MobileNetV2 w/o Self-supervision	4M	0.827	0.743
MobileNetV2 w/ Self-supervision	4M	0.854	0.761
ConvLSTM	16M	0.879	0.788
ResNet18 w/o Self-supervision	11M	0.912	0.785
ResNet18 w/ Self-supervision	**11M**	**0.934**	**0.866**

Table 2. Comparison of ActSonic performance under different deep learning encoders and training strategies. The number of trainable parameters for each model is reported in millions (M).

We developed a self-supervised deep learning pipeline for ActSonic, utilizing ResNet18 [21] as the backbone encoder. The performance of this network is compared with architectures having different encoders and training strategies in Table 2. We present the performance of MobileNetV2 [79] and the ConvLSTM architecture (ResNet18 encoder followed by an LSTM decoder with two layers) in Table 2. We also evaluate the impact of self-supervised pretraining on convolutional encoders (ResNet18 and MobileNetV2). Additionally, the number of trainable parameters (in millions) for each model is reported in Table 2.

Observing Table 2, we note that while self-supervision doesn't exhibit significant performance improvement in the controlled sessions, it does demonstrate an impact in maintaining performance in variable unconstrained scenarios. Furthermore, in comparison to the number of parameters of MobileNetV2, ResNet18 has a larger memory footprint. This observation is particularly valuable in the scenarios involving the performance-inference time tradeoff of the ActSonic system. Additionally, ConvLSTM exhibits worse performance compared to self-supervised ResNet18 despite having the explicit capacity to model temporal dependency.

7.4 Comparison of Performance with Prior Systems

System	Sensing Modality	Device Type	Number of Activities	Activity Examples	Performance (Accuracy)	Study Design
BodyScope [101]	Passive Acoustics	Bluetooth headset	12	eating, drinking, laughing, coughing	79.50%	Small-scale, In-the-wild, 4 activities
Ubicoustics [37]	Passive Acoustics	Commodity electronic devices with mic	30	chopping, baby crying, knocking, speech, alarm clock, etc.	89.60%	In-the-wild
PrivacyMic [25]	Passive Ultrasonic and Infrasonic sensing	Customized hardware board	10	mixer, microwave, kitchen sink, shredder, toilet, etc.	95%	Homes and commercial buildings
SAMoSA [66]	IMU and Subsampled Passive Audio	Smartwatch	26	drill, blender, microwave, coughing, toothbrushing, etc.	92.20%	At participants' home, activities performed according to instruction
DiffAct [55]	Camera	Body-mounted egocentric	71	take cup, preparing coffee, etc.	82.20%	Researcher data
ActSonic	**Active Acoustic Sensing**	**Commodity Eyeglasses**	**27**	**toothbrushing, flossing, eating, drinking, washing hands, coughing, reading book, wiping surface, etc.**	**93.40%**	**Semi-in-the-wild, at participants' home, naturalistic setting**

Table 3. Comparison of performance of activity recognition systems with similar sensing modality. Note that the systems were not evaluated on the same dataset. Therefore, numerical differences may not provide a fair comparison. User study evaluation information is also provided in the table.

We compare ActSonic with other activity recognition systems in Table 3. ActSonic, an active acoustic sensing-based activity recognition system, is compared with systems utilizing audible acoustic data, passively sensed inaudible (ultrasonic and infrasonic) acoustic data, motion data (captured through IMU), and egocentric camera data.

Activity recognition systems rely on passively sensed audible signals (such as BodyScope [101], Ubicoustics [37]) or inaudible acoustics (PrivacyMic [25]). These systems operate on the premise that different actions generate distinct acoustic signatures in the environment. Consequently, their activity sets include actions like chopping, baby crying, or using a mixer. However, many daily activities, such as wearing clothes, reading, or flossing, do not produce discernible sounds suitable for activity inference. Moreover, these activities may not necessarily involve the user wearing the device, especially with wearables. Additionally, previous studies have mainly focused on smaller-scale experiments with specialized sensors or large-scale video-only benchmarks (e.g., DiffAct [55]), with few considering real-world settings.

On the other hand, ActSonic can track fine-grained actions such as flossing or wiping surfaces since its sensing mechanism relies on tracking the movement of different body parts simultaneously. It is evident from Table 3 that ActSonic achieves higher recognition performance (Macro F1-score over 90%) in naturalistic environments, with an extensive set of fine-grained everyday activities. ActSonic's utilization of an active acoustic sensing mechanism enables the capture of signals representing fine-grained movements that passive sensing-based systems may miss. For instance, it can recognize actions occurring outside the frame of an egocentric camera or actions with minimal audio cues, which passive acoustic sensing may struggle to detect accurately. Additionally, passive acoustic sensing-based methods are vulnerable to changes in environmental parameters, hindering their ability to generalize signals across different environments and making deployment in the wild challenging.

8 DISCUSSION

8.1 Visual Explanation of Learned Parameters of ActSonic Model through Saliency Analysis

Our user study demonstrated competitive accuracy in recognizing 27 everyday activities by learning patterns from reflected inaudible acoustic waves around the body. The reflected acoustic signals are formed with complex multipath echoes from both the body and the surrounding environments. Additionally, the environment can occasionally generate acoustic signals with frequency components that overlap with our emitted frequency range. Therefore, received acoustic signals contain rich information about the environment in addition to the body poses.

A common critique of many data-driven sensing systems is that machine learning algorithms operate as a black box, making it difficult to understand what and how they learn. Given the complex reflections forming the received acoustic signals, it is critically important to understand what the deep learning algorithm actually learns from these signals.

To address this, we conducted saliency analysis using Gradient-weighted Class Activation Mapping (Grad-CAM) [81] to visualize the ResNet18 encoder's final convolutional layer. Overlaying feature heatmaps on acoustic flow or differential echo profile channels confirms the model's ability to capture the movements of body parts and areas related to each activity. Figure 12 demonstrates Grad-CAM [81] for all activities in the dataset.

As shown in Figure 12, various activities activate distinct regions within the echo profile sliding window in the neural networks. The y-axis represents the distance from reflected echoes to the sensor, while the x-axis denotes the timeline. To help comprehension, we aligned the y-axis with a human body model. Colors in the figure indicate body movement intensity, with red indicating high intensity and dark blue indicating lower intensity.

The visualization clearly shows that the deep learning algorithms allocate significantly higher attention to areas with heavy body movements (e.g., face, chest, arms) during different activities. For instance, in the activity of brushing teeth, repetitive movement near the face region is evident, with the ResNet18 encoder displaying higher gradient values (indicative of heightened attention) in that area around the face (5 cm-17.15 cm) to infer the activity.

While some activities share similar arm movement patterns, fine details in reflected acoustic waves around the face differentiate them. For instance, both drinking and eating involves moving the arm toward the mouth. However, after feeding during eating, the hand usually moves downward, leaving only mouth movements, visible as subtle highlights under 10 cm. Conversely, during drinking, the hand and container (e.g., cup) remain around the mouth, evident as a large area of heated sections (15-30 cm). Additionally, the activity of opening a door demonstrates hand movement to unlock the door, followed by movement to enter the room.

This visualization confirms that our self-supervised model focuses on different explainable regions of the echo profile within the input sliding window to infer everyday activities. It underscores the importance of capturing detailed movements in both the upper body and face areas to distinguish fine-grained activities such as eating versus drinking, which may not be distinguished in isolation.

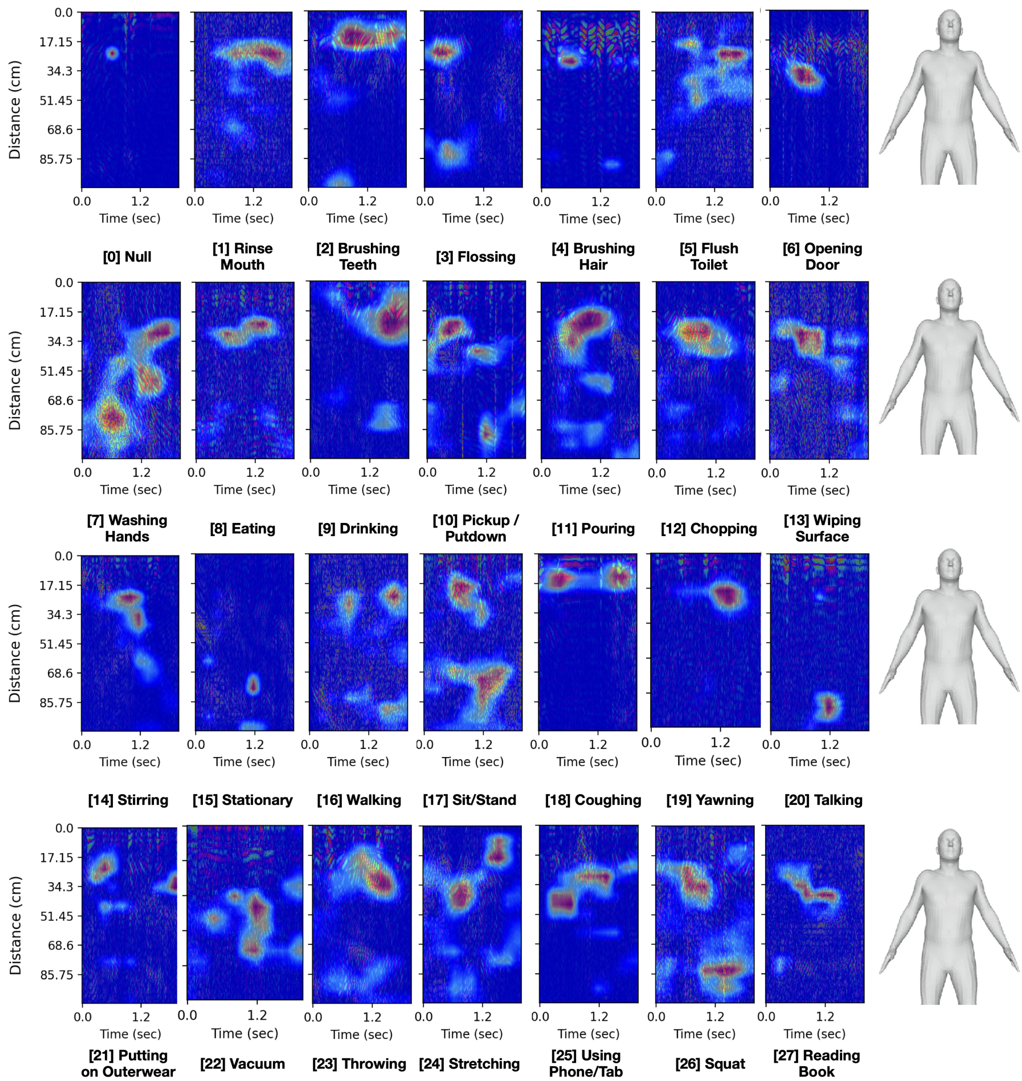

Fig. 12. Selected Grad-CAM [81] heatmaps overlaid on differential echo profiles. As we have a 4-channel input and Grad-CAM aggregates heat maps by channel, we overlay the same (smoothed) heatmap across all four channels. Redder values from the smooth interpolation correspond to higher significance towards class prediction, while bluer values indicate lower significance. We display a SMPL [57] mesh of the body region covered by the sensing range of ActSonic's sliding window input in the rightmost column.

8.2 Deploying ActSonic on Other Wearable Form Factors

ActSonic employs an active acoustic sensing system integrated into eyeglasses to emit inaudible acoustic signals around the body and capture the reflected waves. Due to the low-power and compact nature of the microphones and speakers, this technology can be easily adapted to other wearable form factors for tracking human activities using a similar sensing principle. Here, we evaluate the feasibility of implementing ActSonic on different form factors. Two critical considerations for applying ActSonic to wearable form factors are: 1) the ability to continuously emit acoustic waves around both sides of the upper body and head, and 2) the stability of the wearable form factor during various activities.

Based on these criteria, we anticipate that ActSonic can be seamlessly deployed with minimal modifications on wearable form factors worn around the head (pending training data collection for each form factor), such as necklaces, earbuds, or the recently introduced Humane AI Pin [23] worn on the chest. These form factors provide similar stability to eyeglasses while continuously covering both the upper body and head to track detailed facial and upper body movements.

Wearable form factors attached to limbs, such as wristbands, armbands, or rings, pose greater challenges for applying ActSonic for activity recognition with a single device. Limbs exhibit a wide range of movements, complicating signal interpretation as the source moves concurrently with other body parts. Additionally, signals emitted from one side of the limb may be obstructed by the body, hindering sensing on both sides. Moreover, capturing facial movements from limb-mounted devices is challenging due to the relatively long distance and potential blockages. These issues may be mitigated with updated algorithms and hardware designs, which we plan to explore in future research.

8.3 Model Quantization and MCU Inference

To explore the feasibility of deploying our framework on glasses, we implemented the model pipeline on the MAX78002 microcontroller unit (MCU), leveraging its built-in ultra-low-power CNN accelerator. Initially, we quantized the model by converting high-precision floating-point model parameters to 8-bit integers, a necessary step for deployment on the MAX78002 MCU. Subsequently, we generated a C program for the quantized model inference using the ai8x [2] library provided by the MCU manufacturer. Due to hardware constraints, certain adjustments were made to the model pipeline for compatibility. Notably, 2D convolution kernel sizes were limited to (1×1) or (3×3), with fixed stride size at (1×1). Additionally, the fully connected layer was capped at a maximum of 1024 input neurons on the chip.

Although we successfully ran the ActSonic model on the MAX78002 MCU, the computation of the echo profile and inference of one sliding window through the quantized ResNet18 model on the MCU resulted in a mean latency of 1085.54 milliseconds or 1.09 seconds, yielding a refresh rate of 0.92 Hz. Since we predict activity labels for each 2-second sliding window with 50% overlap, the resulting refresh rate is slower than what is required for real-time inference. The major bottleneck in this process is the computation of the echo profile from the received signal, as this process is not a neural network operation supported by the MAX78002 MCU accelerator. This issue can be alleviated by transforming the cross-correlation operation into an MCU-compatible and accelerated convolution operation, which we leave for future work.

8.4 Robustness to Ambient Noise

The ActSonic system utilizes active acoustic sensing to detect daily activities by monitoring body motions. We evaluated its resilience to environmental noise across 19 participants' homes during the studies. Despite varying environmental factors such as HVAC, running water, TV, and ambient sounds (as detailed in Sec. 5), ActSonic maintains consistent performance independent of environmental settings. This resilience stems from its reliance on ultrasonic frequencies (18 kHz to 24.5 kHz), which exceed most environmental noise sources (recorded at frequencies below 7.5 kHz).

To confirm this hypothesis, we conducted a noise injection experiment. In this experiment, we overlaid noise from five sources commonly found in households onto the acoustic data recorded with ActSonic's sensing module. These noise sources are people talking (61.1 dB), TV/radio (68.8 dB), music (71.5 dB), pets (dogs barking at 82.1 dB and cats meowing at 51.0 dB), and kitchen appliances (microwave at 60.5 dB and blender at 88.7 dB). We then computed the echo profile (as described in Sec. 3.2) and fed it into the ActSonic deep learning framework to evaluate the impact of environmental noise.

To evaluate the impact of different noises on the results, we conducted a one-way repeated measures ANOVA. The F-ratio was computed as 0.526, with degrees of freedom between groups being 5 and degrees of freedom within groups being 108. The p-value was measured as $0.13 > 0.05$, indicating no statistically significant difference between raw and noise-injected acoustic data for tracking everyday activities using the ActSonic system. This lack of significant difference can be attributed to the bandpass filter in the ActSonic pipeline, which cancels out the audible frequency range where the injected noises occur. Nevertheless, further investigation into the impact of ambient noise is crucial to evaluate performance under real-world usage scenarios, which we leave for future work.

8.5 Effectiveness under Multi-user Scenario

Although ActSonic is a wearable system designed to track the everyday activities of the person wearing the eyeglasses, the system can be influenced by the movements of other people and objects surrounding the user, as it relies on the reflected acoustic signals of the transmitted wave. The input to the ActSonic deep learning pipeline is restricted to a sensing range of 101.185 cm, and theoretically, it should not contain any information from reflected signals beyond this range.

In the 19-person user study conducted at participants' homes to evaluate the efficacy of the ActSonic activity recognition framework, we did not impose any restrictions on the number of other people who could be present at home during the study. To further evaluate the multi-user scenario, we conducted a study with 3 participants from the research team. In this evaluation, each person wore the device and performed 21 activities from the ActSonic tracking set (excluding bathroom activities, detailed in Sec. 5.1). While one person performed the activities, the other two were present in the same room, both moving around and staying in place randomly. Additionally, a pedestal fan with a diameter of 16 inches and a height of 47 inches was randomly turned on and off in the same room.

Our leave-one-participant-out evaluation on these 21 activities yielded a mean macro F1 score of 0.857 ± 0.029 across the three users, which is close to the accuracy of the ActSonic system evaluated in the 19-person user study. However, we did not evaluate the scenario where multiple users in the same room each wore a ActSonic prototype. We leave this evaluation for future investigation.

8.6 Health Implication and Usability

ActSonic emits FMCW-encoded ultrasonic waves for active acoustic sensing. To assess health implications, we measured the transmitted signal intensity using a CDC-provided mobile app [7]. The resultant intensity is 68 dB(A), well below the 85 dB limit set by NIOSH [69]. Research [67] on MHz-range ultrasonic exposure suggests muscle tissue discomfort. However, ActSonic operates in the KHz range just above the audible threshold, with no reported issues in this range.

A major concern of active acoustic sensing systems is frequency leakage into the audible range, potentially creating noise from the transmitter. To evaluate this, we collected feedback from participants through a questionnaire after the study. All 19 participants reported that they did not hear any noise or audible sound from the sensing system while wearing the ActSonic form factor. Additionally, none reported any comfort issues. However, two participants mentioned that the glass frame size did not fit their faces, four reported needing their prescription lenses while performing the activities, and one reported that the chest-mounted camera interrupted activities such as eating and reading.

Future investigation should focus on building an attachable sensing module that can be used with any glass frame and integrating the ground truth collection module as an egocentric camera into the eyeglasses form factor.

Future studies will also explore potential audibility among animals and children despite its inaudibility to adults for long-term usage.

8.7 Privacy Preservation

ActSonic utilizes ultrasonic frequency range (18 KHz to 24.5 KHz) to transmit and receive signal. As mentioned in the description of the sensing system in Sec. 3, we apply a bandpass filter on the audio received by the microphone to ensure that ActSonic does not access the audible frequency range to infer activities. Since ActSonic does not require any passively sensed audible acoustic signal, the system does not compromise user privacy by processing sensitive conversation information. Furthermore, the potential of adopting a customized ultrasonic speaker and microphone can further remove the possibility of collecting audible sound.

8.8 Potential Real-world Application

The promising performance of ActSonic recognizing 27 activities in the wild using low-power and minimally obtrusive glasses will significantly lower the barriers to logging everyday activities. It would further create opportunities for many downstream applications that are based on tracking one or multiple types of activities. Here we list a few sample applications :

8.8.1 High-resolution Behavior Data in the Wild for Health Monitoring. ActSonic can be used to journal various everyday activities for different purposes. For instance, journaling food intake behavior is crucial in combating eating disorders, often recorded manually. Previous systems required multiple sensors, had low time resolution (e.g., recognizing a meal every 10 minutes) [84], or needed training data from a user. In contrast, ActSonic can recognize eating moments at 1 Hz with over 90% F1 score in real-world settings without needing training data from a new user, facilitating immediate deployment for eating journaling practices.

Additionally, ActSonic can track 27 everyday activities, many related to health behaviors. Automatically logging these activities can help researchers and clinicians better understand a user's activities in the wild. For instance, eating and drinking behavior can be analyzed for eating disorders and hydration levels, while activities like brushing teeth, flossing, and rinsing the mouth can be tracked with low error rates, aiding dentists in monitoring patient dental behavior.

8.8.2 Tracking Other Activities. In our study, we were able to track 27 activities. However, our system has the potential to recognize other activities that involve movements on the upper body and face. Researchers can potentially replicate our system and customize the frameworks to detect the activity of their interest. For instance, this system can be easily used to automatically track and log the duration and types of the user's exercise routines.

8.9 Limitations and Future Work

Our method is currently limited in the following ways, which we aim to further explore in the future:

8.9.1 Scope of Activity Information in the Dataset. ActSonic recognizes 27 distinct everyday activities within its dataset. However, certain activities in the dataset exhibit variability in execution. For example, actions like yawning or pouring can vary based on contextual factors, which can affect system performance in real-world settings. Addressing this diversity might benefit from a larger dataset and a foundational deep learning feature extractor. Moreover, the activity recognition pipeline of ActSonic lacks external contextual information. Integrating GPS or motion data for fine-grained head or hand movements could aid in understanding environmental affordances and improving activity detection.

8.9.2 Usage of Differential Echo Profile Only. Our system relies solely on the differential echo profile, which may miss static activities with consistent poses. While incorporating the original echo profile might address this, our pilot studies revealed reduced performance and user-dependent features, whereas the differential profile remained more user-independent.

8.9.3 *Multi-label or Concurrent Activity Detection.*

Real-world scenarios involve concurrent activities, a challenge yet to be explored in wearable technology. We aim to explore multi-label classifiers leveraging our system's superior performance.

8.9.4 *Reducing Classification Error during Transitions.*

Many of the misclassification errors occurred when the participants were in transition between two activities, as our system makes predictions every second. In the future, these errors can be easily optimized by developing a state machine, or as simple as a majority-vote mechanism.

9 CONCLUSION

This paper introduces ActSonic, a low-power and unobtrusive action recognition system that employs acoustic sensing on smart glasses. Extensive experiments involving 19 participants in real-world settings showcase ActSonic's adeptness in distinguishing a diverse range of everyday actions across different environments, achieving a mean accuracy of 86.6% in unconstrained user-independent evaluations. We envision ActSonic as a straightforward and efficient supplementary modality for egocentric action recognition, addressing concerns regarding privacy.

ACKNOWLEDGMENTS

This project was supported by the National Science Foundation Grant No. 2239569 and partially by the Cornell University IGNITE Innovation Acceleration Program. ChatGPT was used to polish the writing of the paper.

References

[1] Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. 2021. Attend and Discriminate: Beyond the State-of-the-Art for Human Activity Recognition Using Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 1 (mar 2021), 22 pages. https://doi.org/10.1145/3448083

[2] Analog Devices AI. [n. d.]. ADI MAX78000/MAX78002 Model Training and Synthesis. https://github.com/MaximIntegratedAI/ai8x-synthesis. [Online; accessed 29-Nov-2023].

[3] Amazon.com. [n. d.]. SanDisk 32GB Extreme microSDHC UHS-I Memory Card with Adapter - Up to 100MB/s, C10, U3, V30, 4K, A1, Micro SD - SDSQXAF-032G-GN6MA. https://www.amazon.com/SanDisk-Extreme-microSDHC-UHS-3-SDSQXAF-032G-GN6MA/dp/B06XWMQ81P?th=1. [Online; accessed 29-Nov-2023].

[4] Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*. Springer, 1–17.

[5] Francine Chen, John Adcock, and Shruti Krishnagiri. 2008. Audio privacy: reducing speech intelligibility while preserving environmental sounds. In *Proceedings of the 16th ACM international conference on Multimedia*. 733–736.

[6] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-mounted Miniature Cameras. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 112–125. https://doi.org/10.1145/3379337.3415879

[7] PE Chucri A. Kardous, MS and Ph.D. Peter B. Shaw. [n. d.]. CDC: So How Accurate Are These Smartphone Sound Measurement Apps? https://blogs.cdc.gov/niosh-science-blog/2014/04/09/sound-apps/. [Online; accessed 29-Nov-2023].

[8] Android Developers. [n. d.]. Profile your app performance. https://developer.android.com/studio/profile. [Online; accessed 27-Nov-2023].

[9] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2969–2978.

[10] Peter F Edemekong, Deb Bomgaars, Sukesh Sukumaran, and Shoshana B Levy. 2019. Activities of daily living. (2019).

[11] Xiaoran Fan, David Pearl, Richard Howard, Longfei Shangguan, and Trausti Thormundsson. 2023. Apg: Audioplethysmography for cardiac monitoring in hearables. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[12] Yazan Abu Farha and Jurgen Gall. 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3575–3584.

[13] Jon E Froehlich, Eric Larson, Tim Campbell, Conor Haggerty, James Fogarty, and Shwetak N Patel. 2009. HydroSense: infrastructure-mediated single-point sensing of whole-home water activity. In *Proceedings of the 11th international conference on Ubiquitous computing*. 235–244.

[14] GoPro. 2020. HERO9 Black. https://gopro.com/en/us/shop/cameras/hero9-black/CHDHX-901-master.html. [Online; accessed 12-September-2023].

[15] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 2 (2017), 1–28.

[16] Sidhant Gupta, Matthew S Reynolds, and Shwetak N Patel. 2010. ElectriSense: single-point sensing using EMI for electrical event detection and classification in the home. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 139–148.

[17] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880* (2016).

[18] Harish Haresamudram, David V Anderson, and Thomas Plötz. 2019. On the role of features in human activity recognition. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 78–88.

[19] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. 2020. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 45–49.

[20] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2021. Contrastive Predictive Coding for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 65 (jun 2021), 26 pages. https://doi.org/10.1145/3463506

[21] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 770–778.

[22] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2016. Connectionist temporal modeling for weakly supervised action labeling. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 137–153.

[23] Humane. [n. d.]. Ai Pin - Wearable Ai. https://humane.com/. [Online; accessed 01-May-2024].

[24] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 448–456. http://jmlr.org/proceedings/papers/v37/ioffe15.pdf

[25] 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*. 1–13.

[26] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture sensing using on-body acoustic interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[27] Yincheng Jin, Seokmin Choi, Yang Gao, Jiyang Li, Zhengxiong Li, and Zhanpeng Jin. 2023. TransASL: A Smart Glass based Comprehensive ASL Recognizer in Daily Life. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (<conf-loc>, <city>Sydney</city>, <state>NSW</state>, <country>Australia</country>, </conf-loc>) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 802–818. https://doi.org/10.1145/3581641.3584071

[28] Yincheng Jin, Yang Gao, Yanjun Zhu, Wei Wang, Jiyang Li, Seokmin Choi, Zhangyu Li, Jagmohan Chauhan, Anind K. Dey, and Zhanpeng Jin. 2021. SonicASL: An Acoustic-based Sign Language Gesture Recognizer Using Earphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 67 (jun 2021), 30 pages. https://doi.org/10.1145/3463519

[29] Mone Kijima, Yuta Miyagaw, Hayato Oshita, Norihisa Segawa, Masato Yazawa, and Masa-yuki Yamamoto. 2018. Multiple door opening/closing detection system using infrasound sensor. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 126–127.

[30] Dongkeun Kim, Jinsung Lee, Minsu Cho, and Suha Kwak. 2022. Detector-free weakly supervised group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20083–20093.

[31] Dennis Kimm and David V Thiel. 2015. Hand speed measurements in boxing. *Procedia Engineering* 112 (2015), 502–506.

[32] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[33] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. 2019. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12066–12074.

[34] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. 2011. Activity Recognition Using Cell Phone Accelerometers. *SIGKDD Explor. Newsl.* 12, 2 (mar 2011), 74–82. https://doi.org/10.1145/1964897.1964918

[35] Hyeokhyen Kwon, Gregory D Abowd, and Thomas Plötz. 2018. Adding structural characteristics to distribution-based accelerometer representations for activity recognition using wearables. In *Proceedings of the 2018 ACM international symposium on wearable computers*. 72–75.

[36] Nicholas D Lane, Petko Georgiev, and Lorena Qendro. 2015. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 283–294.

[37] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 213–224.

[38] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300568

[39] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 321–333. https://doi.org/10.1145/2984511.2984582

[40] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic sensors: Towards general-purpose sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3986–3999.

[41] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 156–165.

[42] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Deng, Ke Li, et al. 2024. EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband. *arXiv preprint arXiv:2401.17409* (2024).

[43] Boning Li and Akane Sano. 2020. Extraction and Interpretation of Deep Autoencoder-Based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 49 (jun 2020), 26 pages. https://doi.org/10.1145/3397318

[44] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. LASense: Pushing the Limits of Fine-grained Activity Sensing Using Acoustic Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 21 (mar 2022), 27 pages. https://doi.org/10.1145/3517253

[45] Ke Li, Ruidong Zhang, Boao Chen, Siyuan Chen, Sicheng Yin, Saif Mahmud, Qikang Liang, François Guimbretière, and Cheng Zhang. 2024. GazeTrak: Exploring Acoustic-based Eye Tracking on a Glass Frame. *arXiv preprint arXiv:2402.14634* (2024).

[46] Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen, Mose Sakashita, François Guimbretière, and Cheng Zhang. 2024. EyeEcho: Continuous and Low-power Facial Expression Tracking on Glasses. *arXiv preprint arXiv:2402.12388* (2024).

[47] Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen, Mose Sakashita, François Guimbretière, and Cheng Zhang. 2024. EyeEcho: Continuous and Low-power Facial Expression Tracking on Glasses. *arXiv preprint arXiv:2402.12388* (2024).

[48] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-Power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. 6, 2, Article 62 (jul 2022), 24 pages. https://doi.org/10.1145/3534621

[49] Dawei Liang, Wenting Song, and Edison Thomaz. 2020. Characterizing the Effect of Audio Degradation on Privacy Perception And Inference Performance in Audio-Based Human Activity Recognition. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 32, 10 pages. https://doi.org/10.1145/3379503.3403551

[50] Dawei Liang and Edison Thomaz. 2019. Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–18.

[51] Hyunchul Lim, Guilin Hu, Richard Jin, Hao Chen, Ryan Mao, Ruidong Zhang, and Cheng Zhang. 2023. C-Auth: Exploring the Feasibility of Using Egocentric View of Face Contour for User Authentication on Glasses. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*. 6–10.

[52] Hyunchul Lim, Yaxuan Li, Matthew Dressa, Fang Hu, Jae Hoon Kim, Ruidong Zhang, and Cheng Zhang. 2022. BodyTrak: Inferring Full-Body Poses from Body Silhouettes Using a Miniature Camera on a Wristband. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 154 (sep 2022), 21 pages. https://doi.org/10.1145/3552312

[53] Hyunchul Lim, Ruidong Zhang, Samhita Pendyal, Jeyeon Jo, and Cheng Zhang. 2023. D-Touch: Recognizing and Predicting Fine-grained Hand-face Touching Activities Using a Neck-mounted Wearable. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (<conf-loc>, <city>Sydney</city>, <state>NSW</state>, <country>Australia</country>, </conf-loc>) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 569–583. https://doi.org/10.1145/3581641.3584063

[54] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[55] Daochang Liu, Qiyue Li, AnhDung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. 2023. Diffusion Action Segmentation. *arXiv preprint arXiv:2303.17959* (2023).

[56] Jialin Liu, Dong Li, Lei Wang, Fusang Zhang, and Jie Xiong. 2022. Enabling Contact-free Acoustic Sensing under Device Motion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 128 (sep 2022), 27 pages. https://doi.org/10.1145/3550329

[57] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2023. *SMPL: A Skinned Multi-Person Linear Model* (1 ed.). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3596711.3596800

[58] LowPowerLab. [n. d.]. Current Ranger. https://lowpowerlab.com/guide/currentranger/. [Online; accessed 29-Nov-2023].

[59] Shihan Lu and Heather Culbertson. 2023. Active Acoustic Sensing for Robot Manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 3161–3168. https://doi.org/10.1109/IROS55552.2023.10342481

[60] Paul Lukowicz, Jamie A Ward, Holger Junker, Mathias Stäger, Gerhard Tröster, Amin Atrash, and Thad Starner. 2004. Recognizing workshop activity using body worn microphones and accelerometers. In *Pervasive Computing: Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004. Proceedings 2*. Springer, 18–32.

[61] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: Multi-level attention mechanism for multimodal human activity recognition.. In *IJCAI*. 3109–3115.

[62] Saif Mahmud, Ke Li, Guilin Hu, Hao Chen, Richard Jin, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2023. PoseSonic: 3D Upper Body Pose Estimation Through Egocentric Acoustic Sensing on Smartglasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 111 (sep 2023), 28 pages. https://doi.org/10.1145/3610895

[63] Saif Mahmud, M. T. H. Tonmoy, Kishor Kumar Bhaumik, A. M. Rahman, M. A. Amin, M. Shoyaib, Muhammad Asif Hossain Khan, and A. Ali. 2020. Human Activity Recognition from Wearable Sensor Data Using Self-Attention. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain*.

[64] maxim integrated. [n. d.]. MAX98357A/ MAX98357B Tiny, Low-Cost, PCM Class D Amplifier with Class AB Performance. https://www.analog.com/media/en/technical-documentation/data-sheets/MAX98357A-MAX98357B.pdf. [Online; accessed 29-Nov-2023].

[65] Johannes Meyer, Adrian Frank, Thomas Schlebusch, and Enkeljeda Kasneci. 2022. A CNN-Based Human Activity Recognition System Combining a Laser Feedback Interferometry Eye Movement Sensor and an IMU for Context-Aware Smart Glasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 172 (dec 2022), 24 pages. https://doi.org/10.1145/3494998

[66] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–19.

[67] David Baeza Moyano, Daniel Arranz Paraiso, and Roberto Alonso González-Lezcano. 2022. Possible effects on health of ultrasound exposure, risk factors in the work environment and occupational safety review. In *Healthcare*, Vol. 10. MDPI, 423.

[68] Vishvak S Murahari and Thomas Plötz. 2018. On attention models for human activity recognition. In *Proceedings of the 2018 ACM international symposium on wearable computers*. 100–103.

[69] William J Murphy and John R Franks. 2002. Revisiting the NIOSH criteria for a recommended standard: Occupational noise exposure. *The Journal of the Acoustical Society of America* 111, 5 (2002), 2397.

[70] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones. In *Proceedings of the 13th annual international conference on mobile systems, applications, and services*. 45–57.

[71] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. FingerIO: Using Active Sonar for Fine-Grained Finger Tracking. 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, 1515–1525. https://doi.org/10.1145/2858036.2858580

[72] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. 2023. AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12999–13008.

[73] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[74] PJRC. [n. d.]. Teensy® 4.1 Development Board. https://www.pjrc.com/store/teensy41.html. [Online; accessed 29-Nov-2023].

[75] Will Price, Carl Vondrick, and Dima Damen. 2022. Unweavenet: Unweaving activity stories. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13770–13779.

[76] Hangwei Qian, Sinno Jialin Pan, Bingshui Da, and Chunyan Miao. 2019. A Novel Distribution-Embedded Neural Network for Sensor-Based Activity Recognition.. In *IJCAI*, Vol. 2019. 5614–5620.

[77] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal Deep Learning for Activity and Context Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 157 (jan 2018), 27 pages. https://doi.org/10.1145/3161174

[78] Aaqib Saeed, Flora D Salim, Tanir Ozcelebi, and Johan Lukkien. 2020. Federated self-supervised learning of multisensor representations for embedded intelligence. *IEEE Internet of Things Journal* 8, 2 (2020), 1030–1040.

[79] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[80] Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. 2021. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11225–11234.

[81] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (oct 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[82] Nordic Semiconductors. [n. d.]. nRF52840 Multiprotocol Bluetooth 5.4 SoC supporting Bluetooth Low Energy, Bluetooth mesh, NFC, Thread and Zigbee. https://www.nordicsemi.com/products/nrf52840. [Online; accessed 29-Nov-2023].

[83] SGWireless. [n. d.]. SGW111X BLE Modules. https://www.sgwireless.com/product/SGW111X. [Online; accessed 29-Nov-2023].

[84] Jaemin Shin, Seungjoo Lee, Taesik Gong, Hyungjun Yoon, Hyunchul Roh, Andrea Bianchi, and Sung-Ju Lee. 2022. MyDJ: Sensing Food Intakes with an Attachable on Your Eyeglass Frame. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 341, 17 pages. https://doi.org/10.1145/3491102.3502041

[85] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[86] David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. MM-Fit: Multimodal Deep Learning for Automatic Exercise Logging across Sensing Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 168 (dec 2020), 22 pages. https://doi.org/10.1145/3432701

[87] Rujia Sun, Xiaohe Zhou, Benjamin Steeper, Ruidong Zhang, Sicheng Yin, Ke Li, Shengzhang Wu, Sam Tilsen, Francois Guimbretiere, and Cheng Zhang. 2023. EchoNose: Sensing Mouth, Breathing and Tongue Gestures inside Oral Cavity using a Non-contact Nose Interface. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers* (Cancun, Quintana Roo, Mexico) *(ISWC '23)*. Association for Computing Machinery, New York, NY, USA, 22–26. https://doi.org/10.1145/3594738.3611358

[88] Wei Sun, Franklin Mingzhe Li, Congshu Huang, Zhenyu Lei, Benjamin Steeper, Songyun Tao, Feng Tian, and Cheng Zhang. 2021. ThumbTrak: Recognizing micro-finger poses using a ring with proximity sensing. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 1–9.

[89] Wei Sun, Franklin Mingzhe Li, Benjamin Steeper, Songlin Xu, Feng Tian, and Cheng Zhang. 2021. Teethtap: Recognizing discrete teeth gestures using motion and acoustic sensing on an earpiece. In *26th International Conference on Intelligent User Interfaces*. 161–169.

[90] TDK. [n. d.]. ICS-43434 Multi-Mode Microphone with I^2S Digital Output. https://invensense.tdk.com/products/ics-43434/. [Online; accessed 29-Nov-2023].

[91] Catherine Tong, Shyam A Tailor, and Nicholas D Lane. 2020. Are accelerometers for activity recognition a dead-end?. In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*. 39–44.

[92] M. Tanjid Hasan Tonmoy, Saif Mahmud, A. K. M. Mahbubur Rahman, M. Ashraful Amin, and Amin Ahsan Ali. 2021. Hierarchical Self Attention Based Autoencoder for Open-Set Human Activity Recognition. In *Advances in Knowledge Discovery and Data Mining*, Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty (Eds.). Springer International Publishing, Cham, 351–363.

[93] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. 2018. Context Recognition In-the-Wild: Unified Model for Multi-Modal Sensors and Multi-Label Classification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 168 (jan 2018), 22 pages. https://doi.org/10.1145/3161192

[94] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. 1, 4, Article 170 (jan 2018), 20 pages. https://doi.org/10.1145/3161188

[95] Jamie A Ward, Paul Lukowicz, Gerhard Troster, and Thad E Starner. 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE transactions on pattern analysis and machine intelligence* 28, 10 (2006), 1553–1567.

[96] Jason Wu, Chris Harrison, Jeffrey P Bigham, and Gierad Laput. 2020. Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[97] Yadong Xie, Fan Li, Yue Wu, and Yu Wang. 2021. HearFit: Fitness Monitoring on Smart Speakers via Active Acoustic Sensing. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488811

[98] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015).

[99] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior* 37 (2013), 217–230.

[100] Shuochao Yao, Yiran Zhao, Huajie Shao, Dongxin Liu, Shengzhong Liu, Yifan Hao, Ailing Piao, Shaohan Hu, Su Lu, and Tarek F Abdelzaher. 2019. Sadeepsense: Self-attention deep learning framework for heterogeneous on-device sensors in internet of things applications. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1243–1251.

[101] Koji Yatani and Khai N Truong. 2012. Bodyscope: a wearable acoustic sensor for activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 341–350.

[102] Tomohiro Yokota and Tomoko Hashida. 2016. Hand Gesture and On-body Touch Recognition by Active Acoustic Sensing throughout the Human Body. In *Adjunct Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 113–115. https://doi.org/10.1145/2984751.2985721

[103] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott Gilliland, Thomas Ploetz, Thad E Starner, Omer T Inan, and Gregory D Abowd. 2017. FingerSound: Recognizing unistroke thumb gestures using a ring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–19.

[104] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, Omer Inan, and Gregory D. Abowd. 2018. FingerPing: Recognizing Fine-grained Hand Poses using Active Acoustic On-body 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, 1–10. https://doi.org/10.1145/3173574.3174011

[105] Jiahao Zhang, Stephen Gould, and Itzik Ben-Shabat. 2020. Vidat—ANU CVML Video Annotation Tool. https://github.com/anucvml/vidat.

[106] Rui Zhang and Oliver Amft. 2017. Monitoring chewing and eating in free-living using smart eyeglasses. *IEEE journal of biomedical and health informatics* 22, 1 (2017), 23–32.

[107] Rui Zhang and Oliver Amft. 2018. Free-living eating event spotting using EMG-monitoring eyeglasses. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 128–132.

[108] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-Obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 852, 18 pages. https://doi.org/10.1145/3544548.3580801

[109] Ruidong Zhang, Jihai Zhang, Nitish Gade, Peng Cao, Seyun Kim, Junchi Yan, and Cheng Zhang. 2022. EatingTrak: Detecting Fine-Grained Eating Moments in the Wild Using a Wrist-Mounted IMU. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 214 (sep 2022), 22 pages. https://doi.org/10.1145/3546749

[110] Shijia Zhang, Taiting Lu, Hao Zhou, Yilin Liu, Runze Liu, and Mahanth Gowda. 2023. I Am an Earphone and I Can Hear My User's Face: Facial Landmark Tracking Using Smart Earphones. *ACM Trans. Internet Things* 5, 1, Article 1 (dec 2023), 29 pages. https://doi.org/10.1145/3614438