ActSonic: Everyday Activity Recognition on Smart Glasses using Active Acoustic Sensing

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
FRANÇOIS GUIMBRETIÈRE, Cornell University, USA
CHENG ZHANG, Cornell University, USA

In this paper, we introduce ActSonic, an intelligent, low-power active acoustic sensing system integrated into eyeglasses. ActSonic is designed to recognize 27 different everyday activities (e.g., eating, drinking, toothbrushing). It only needs a pair of miniature speakers and microphones mounted on each hinge of eyeglasses to emit ultrasonic waves to create an acoustic aura around the body. Based on the position and motion of various body parts, the acoustic signals are reflected with unique patterns captured by the microphone and analyzed by a customized self-supervised deep learning framework to infer the performed activities. ActSonic was deployed in a user study with 19 participants across 19 households to evaluate its efficacy. Without requiring any training data from a new user (leave-one-participant-out evaluation), ActSonic was able to detect 27 activities with an inference resolution of 1 second, achieving an average F1-score of 86.6% in an unconstrained setting and 93.4% in a prompted setting.

1 INTRODUCTION

Wearables are widely used globally to monitor daily behaviors and activities. Despite advancements in artificial intelligence, accurately tracking basic everyday human activities in real-world settings, often referred to as "in the wild," remains a significant challenge for these devices. For example, common wearables like smartwatches and glasses encounter difficulty in precisely tracking complex actions such as eating or drinking. The primary challenge in recognizing everyday human activities lies in consistently capturing high-quality information regarding different body parts involved in these activities.

Most wearable devices integrate inertial measurement units (IMUs), which, despite being small and low-power, often lack the necessary resolution for recognizing detailed human behaviors [83, 96]. Researchers have explored wearable cameras [50, 67] as an alternative to improve activity identification, showcasing promising performance. However, integrating cameras into wearables faces practical limitations due to high energy consumption and significant data generation, raising privacy concerns in everyday contexts. Another approach utilizes microphones in wearables to capture sounds produced during various activities in the wild, offering low-power and cost-effective options. Yet, this method struggles with activities that do not generate distinct sounds (e.g., exercise, reading) [23, 34, 58], constituting a significant portion of daily activities.

Authors' addresses: 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; François Guimbretière, Cornell University, Ithaca, NY, USA, francois@cs.cornell.edu; Cheng Zhang, Cornell University, Ithaca, NY, USA, chengzhang@cornell.edu.

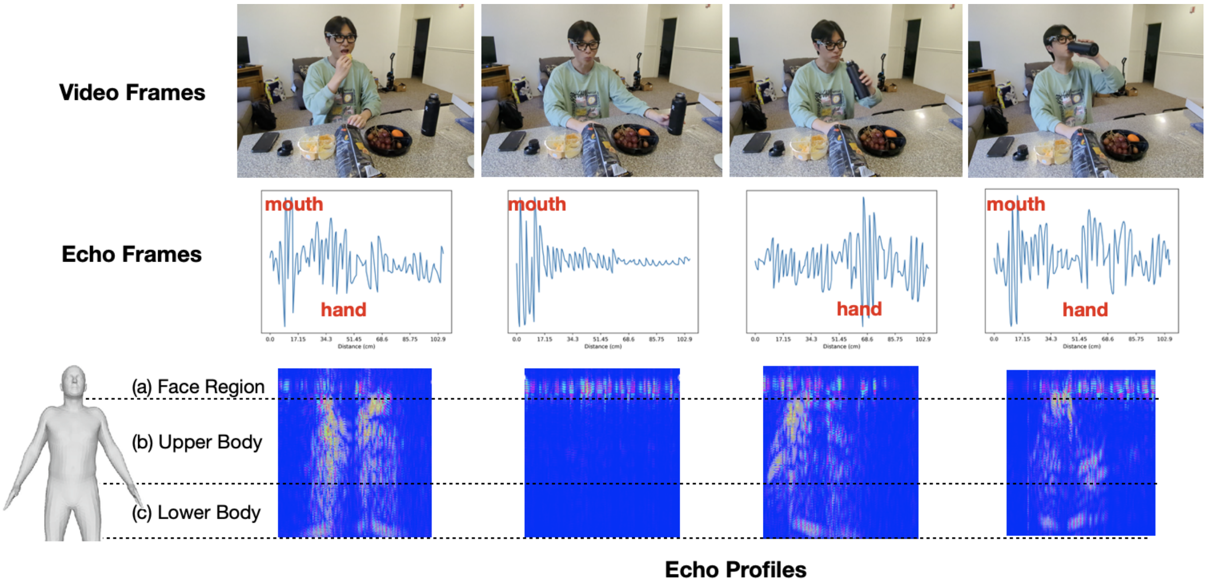


Fig. 1. Overview of the active acoustic sensing principle of ActSonic: The x-axis of the echo frames (in the 2nd row) represents the distance of echo reception. The corresponding video frames (in the first row) serve as activity references. The echo profile, created by stacking multiple echo frames, provides a spatio-temporal 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.

Numerous activities in our daily lives involve the movements of different body parts, especially on the upper body and face. For instance, eating combines hand-to-mouth motion (using a spoon or chopsticks) with chewing. Therefore, activity recognition systems require detailed hand and mouth movement data to track fine-grained episodes of eating. Conversely, rinsing the mouth or yawning has very similar hand movements compared to eating. However, there are different movements of facial muscles if we follow the temporal progression of these activities. It is challenging to track the pose and movements of multiple body parts simultaneously at a single instrumentation point (e.g. IMU). As a result, existing wearable activity recognition systems suffer from a lack of precision and granularity in tracking complex activities in everyday life.

In this paper, we introduce ActSonic, a self-supervised and low-power activity recognition system integrated into eyeglasses, based on active acoustic sensing. ActSonic is the first to demonstrate the feasibility of using active acoustic sensing on a wearable device to recognize 27 types of everyday activities without the need to collect any training data from a new user. Due to the low-power nature of acoustic sensors, it can operate for over 21 hours with a battery capacity equivalent to that of Google Glass (570 mAh). This work is motivated by the research question of creating a wearable sensing platform with a single instrumentation point to track a broad spectrum of everyday activities in real-world settings. It draws inspiration from recent advancements in using active acoustic sensing to monitor facial expressions [42, 44] and upper body postures [54] via glasses or earphones, as these body postures are central to performing a variety of daily activities.

We developed ActSonic by attaching a pair of miniature, low-power, off-the-shelf microphones and speakers to the hinges of glasses, respectively. 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 patterns captured by the microphone. We developed a customized self-supervised deep learning framework to interpret the reflected signals, which are presented with complex multipath echoes and include rich information about movements on both the face and upper body, 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 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 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. The leave-one-participant-out evaluation showed that ActSonic achieved 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 has significantly advanced wearable-based activity recognition, demonstrating promising performance across various dimensions. Unlike previous data-driven systems that require user-specific training data, ActSonic accurately recognizes 27 activities in participants' homes without the need for individualized training data collection. Additionally, it offers an affordable and low-power hardware solution, enabling over 21 hours of continuous operation on wearables with small batteries—a notable departure from high-power signature sensing systems, such as cameras. While few wearable-based activity recognition systems have been evaluated in uncontrolled scenarios, ActSonic demonstrates robust performance in recognizing a diverse set of 27 activities in real-world settings. Moreover, compared to many previous works relying on multiple sensors for limited activities, ActSonic achieves the recognition of 27 activities with a single device, significantly reducing barriers to tracking everyday activities using glasses in real-world contexts.

In summary, the contributions of the paper are:

- The first demonstration of utilizing low-power active acoustic sensing for fine-grained activity recognition.
- We developed a self-supervised deep learning framework that is able to distinguish 27 activities from the received reflected acoustic signals
- We conducted a semi-in-the-wild study and an in-the-wild study with 19 participants at 19 homes to collect over 40 hours of activity data, demonstrating a promising performance with F1 scores of 93.4% and 86.6% respectively

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 [12, 15], 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, heart rate sensors, etc. Early research into IMU-based activity recognition relied on hand-crafted features [4, 17, 32, 68] generated from sensor readings. These methods [31] 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 [14, 16, 55, 60, 65] 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 [35], body gestures [36], etc. These deep learning architectures, incorporating convolutions, recurrence, and transformers [53, 84, 90], 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 [18], signal correspondence learning [70], and contrastive predictive coding [19] 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 [83] 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, 46–48, 64, 67] or systems installed in an environment [9]. In vision-based action segmentation approaches [11, 21, 30, 38, 72], they are tasked with assigning activity labels to each frame of the video. These vision-based approaches adopt weakly supervised [28] or unsupervised [30, 72] or self-supervised [67] modeling techniques to detect activities by learning temporal embedding of the video frames. On the other hand, multimodal approaches [1, 40, 57, 78, 85] 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 [37, 69], body or finger gestures [52, 80, 81, 87, 92, 93]. 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 [33, 34, 45, 88]. 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, recent works [5, 58] adopt subsampling and other preprocessing techniques to make the audio unintelligible and then recognize activities based on that. Additionally, passively sensed audio from inaudible frequency ranges (infrasonic and ultrasonic) has been utilized to recognize daily activities [23, 24, 27]. 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 [43, 44, 79, 95], hand poses [39], sign language gesture [25, 26], body pose [54], gaze [41]and vital signs [62, 86].

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 [42, 44, 54, 95] has underscored the efficacy of active acoustic sensing in detecting facial changes and upper body poses, no existing system has seamlessly integrated

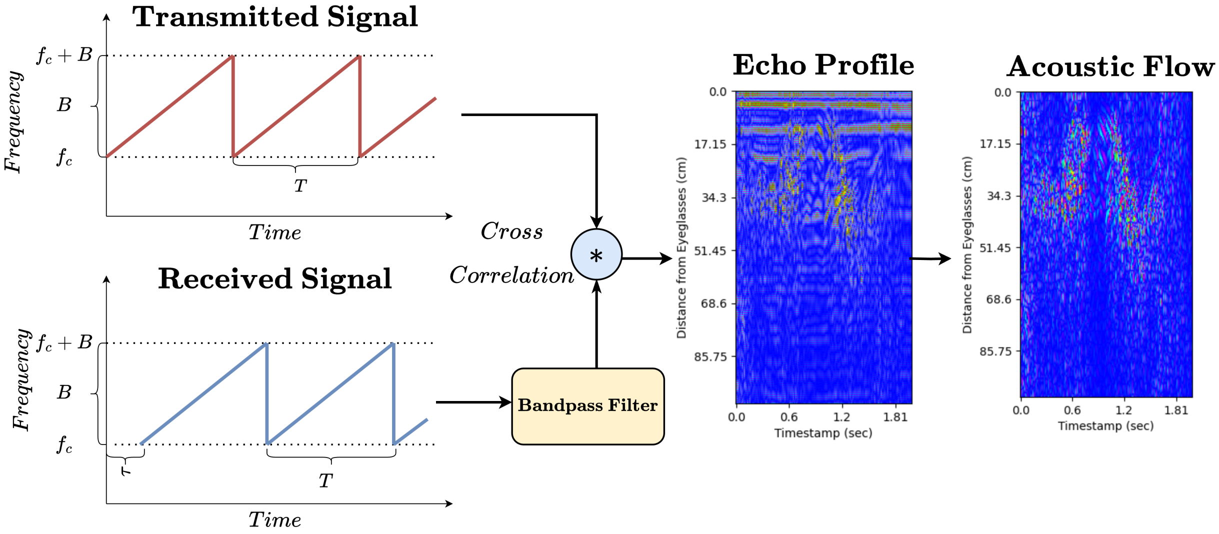


Fig. 2. Overview of echo profile and acoustic flow calculation. For the echo profile, we cross-correlate the transmitted signal with a bandpass filter applied over the received signal (to ensure only specific frequencies are returned). This allows us to capture the direct echo profile, and we can calculate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

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 device, 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.

3.1 Configuration of Active Acoustic Signal

The active acoustic sensing system in ActSonic integrates two pairs of ultrasonic transmitters and receivers on eyeglass hinges. Utilizing Frequency Modulated Continuous Wave (FMCW) chirps, ActSonic emits ultrasonic signals with linearly modulated frequencies ranging from 18 to 21.5 KHz and 21.5 to 24.5 KHz for the left and right transmitters, respectively, each with a bandwidth of $B = 3$ KHz. The ActSonic system samples these FMCW chirps at $f_s = 50$ KHz. By employing cross-correlation-based FMCW [86], the system achieves a minimum discernible distance, or resolution, of $\frac{c}{2f_s} = 0.343$ cm ($c = 343$ m/s in dry air at $20°C$). With a sweep period T of 12 ms for transmitted FMCW chirps (comprising $N = 600$ samples), ActSonic's maximum sensing range extends to approximately 2 meters. 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.

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 [86] 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* [44, 63, 86, 95], 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 $T_{left} \to R_{left}$, $T_{left} \to R_{right}$, $T_{right} \to R_{left}$, and $T_{right} \to R_{right}$. 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

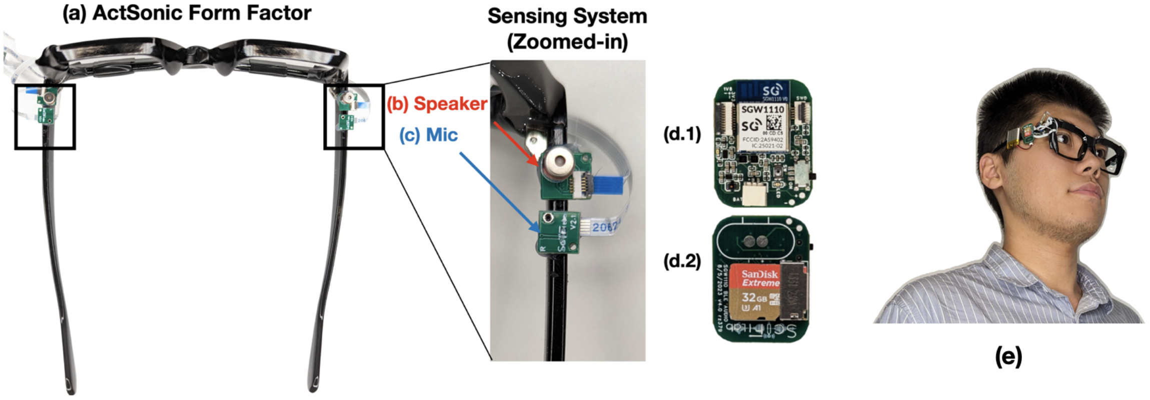


Fig. 3. Hardware of ActSonic: **(a)** Eyeglasses form factor, **(b)** Transmitter or speaker, **(c)** Receiver or microphone (dimension of the sensor board of (b) and (c) is $9mm \times 9mm$), (d) Front (d.1) and back (d.2) of customized PCB board (dimension $18mm \times 23mm$) with low-power nRF52840 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 [82], following a design similar to that shown in [44]. Managed by a Teensy 4.1 microcontroller [66], 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 [56]. 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 microcontroller, powered by an ARM M7 core and exhibiting higher power consumption due to its characteristics. To highlight the power efficiency of our acoustic sensing, we designed a low-power variant featuring an nRF52840 microcontroller [74] (depicted in Figure 3(d)), 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 [75] 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 Fig. 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.

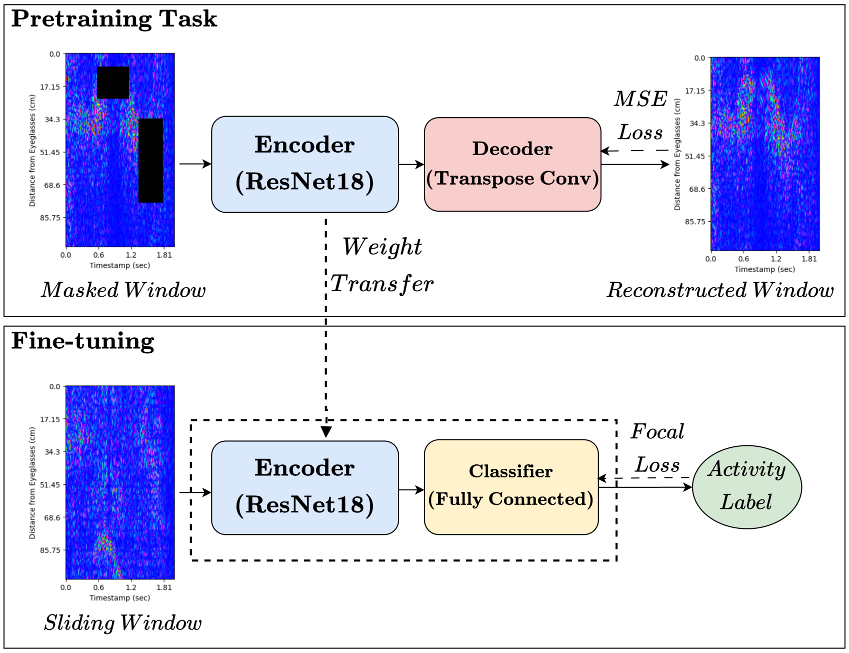


Fig. 4. Deep learning model architecture for ActSonic. Within the self-supervised **pretraining** stage, we mask out specific sections of the input echo profile and train an encoder-decoder architecture to reconstruct the input echo profile (given a lightweight decoder) supervised by an MSE loss. We 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 Fig. 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 Fig. 4, the masked window goes through a ResNet18 [20] 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 Fig. 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 [22], Leaky ReLU activation [89], and dropout [77] 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 [49] 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) = $(4 \times 295 \times 166)$. 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 [29] 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:

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

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 fully in-the-wild study, both conducted in participants' homes in an unconstrained 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, 23, 34, 58] 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, flossing, brush_hair, flush_toilet, open_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/cleaning_floor, throwing, stretching, using_phone/tab, squat, reading_book

5.2 Participants and Study Schedule

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 participants for a fully-in-the-wild study. Among the total 19 participants, with an average age of (24.737 ± 3.445) years, ranging from

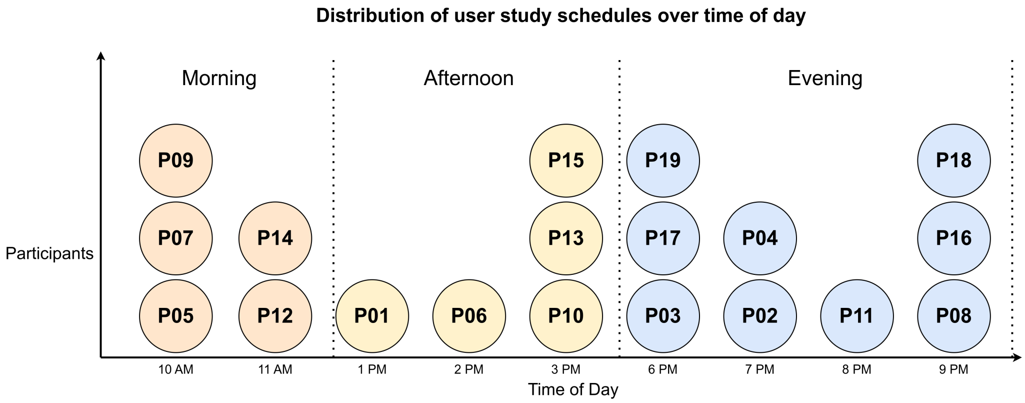


Fig. 5. Distribution of participant schedules for the user study over time, where x-axis represents time and y-axis represents participant count. We split the participants into three general groups ("morning" as 7 am - 12 pm, "afternoon" as 12 pm - 6 pm, and "evening" as 6 pm - 11 pm) and ensure that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

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. As shown in Fig. 5, each point represents the start time of a study session.

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 [13] 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 *in-the-wild 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 ***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.

5.4.2 ***Study - 02:*** *In-the-wild User Study at Participants' Home.*

We designed a longer-duration in-the-wild 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 in-the-wild 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 in-the-wild 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) [94] to annotate all ground-truth egocentric video data with action annotations. Subsequently, we synchronized the timestamps of acoustic and video data using a clapping action with 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[33, 58], which evaluated the activity recognition in similar manner (albeit different activities), ActSonic has significantly better performance with a wider range of activities.

However, we also admitted 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 of how they perform activities in real-world settings.

6.2 Evaluation of User Independent Model on Unconstrained Sessions

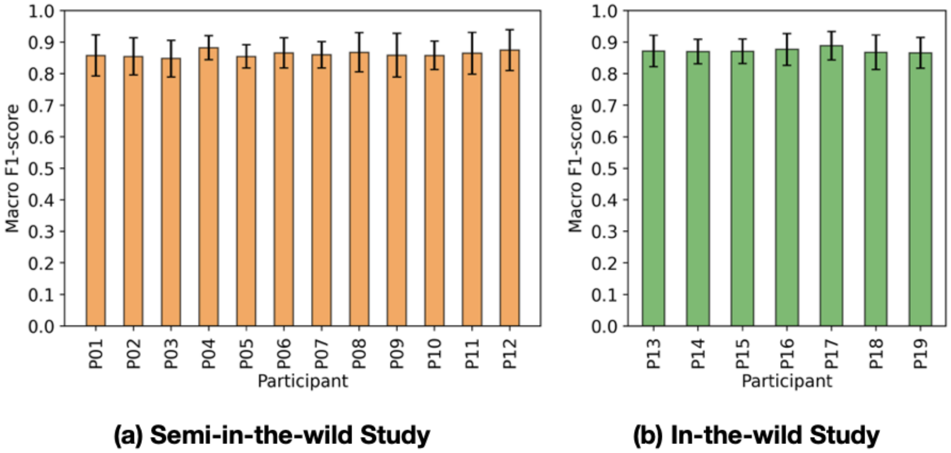


Fig. 6. Leave-one-participant-out performance evaluation of unconstrained sessions

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

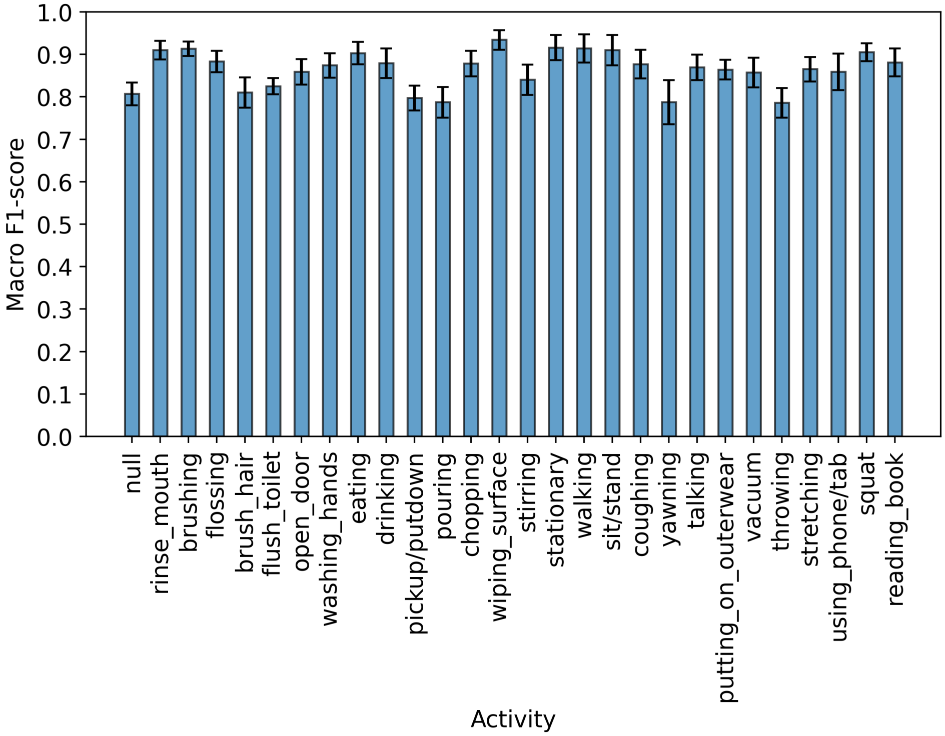


Fig. 7. In-the-wild evaluation of the performance of different activities.

study. Hence, these sessions entail activity samples that might exhibit more diverse motion profiles compared to prompted sessions.

Our evaluation of the unconstrained sessions from both the semi-in-the-wild and in-the-wild 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 in-the-wild study. To assess the unconstrained sessions of users P13 to P19 from the in-the-wild 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 in-the-wild participants respectively. Furthermore, Figure 7 displays the average macro F1-score and standard deviation for each of the in-the-wild 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 1Hz. Although the system's performance in an in-the-wild scenario is comparatively lower than in prompted sessions, it is apparent in Figure 9 that

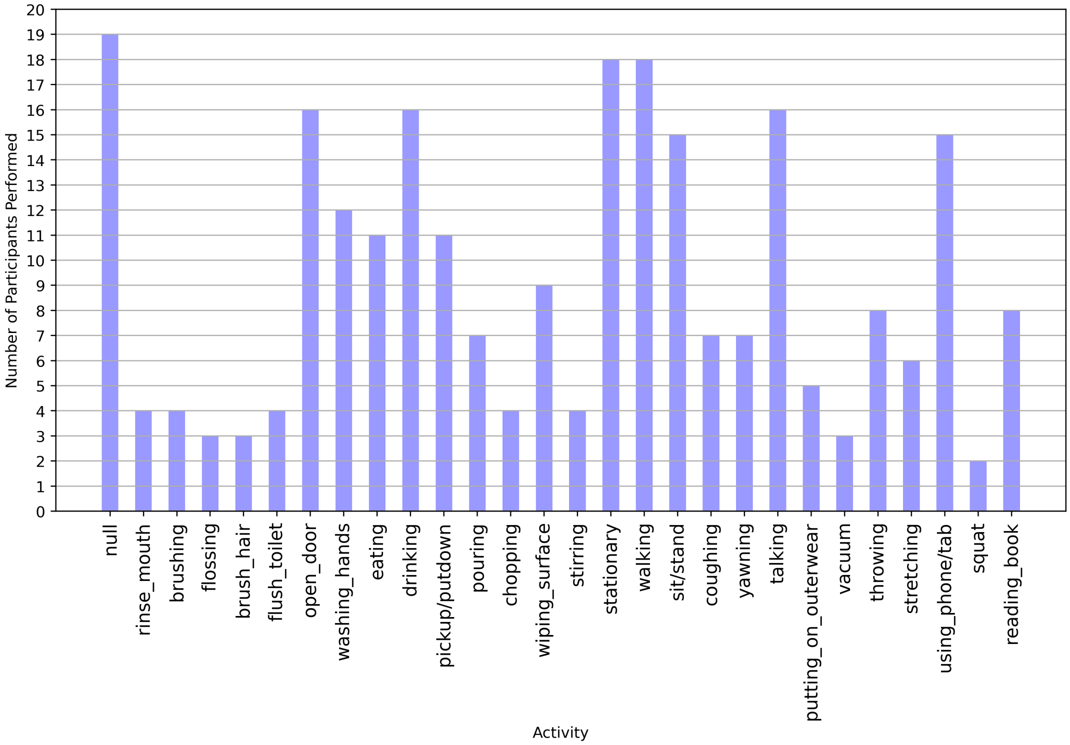


Fig. 8. Activity-wise number of participants for the unconstrained sessions: with activity labels on the x-axis and number of participants on the y-axis.

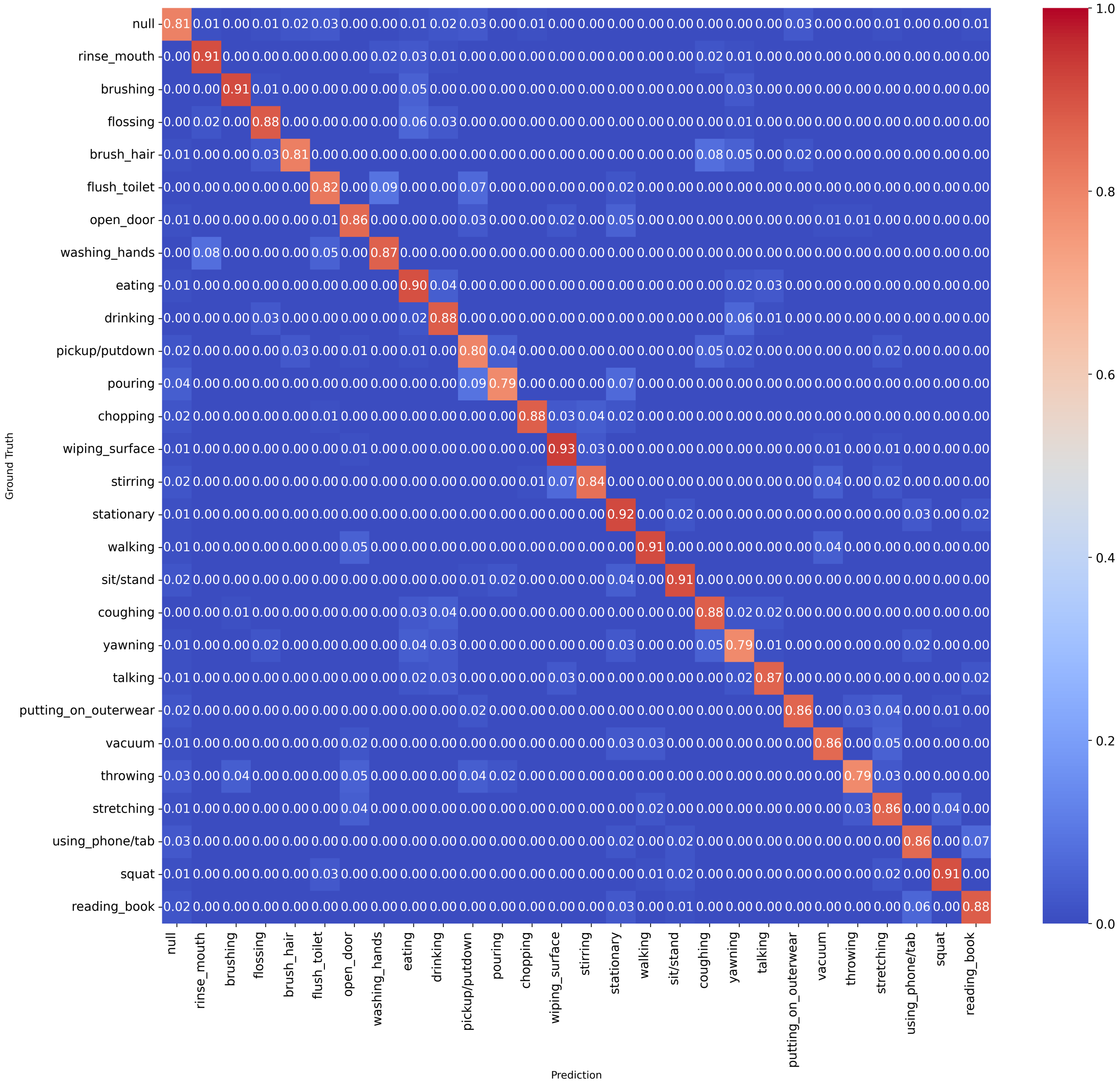
the activity recognition remains accurate, considering the user variability and class imbalance present in the in-the-wild sessions.

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 in-the-wild 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.

The reported performance is a significant advancement in wearable-based activity recognition in the wild, as it is the first wearable-based (non-camera) activity recognition system that was evaluated in such an unconstrained environment, with a wide variety of activities in a user-independent manner, while lasting over 20 hours.

6.3 Power Signature of the System

Our ActSonic system initially consumed 577.8 mW 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 [7], displayed an average operation of 96.5 mW (4.02 V, 24.0 mA) while saving all data to



	null	rinse_mouth	brushing	flossing	brush_hair	flush_toilet	open_door	washing_hands	eating	drinking	pickup/putdown	pouring	chopping	wiping_surface	stirring	stationary	walking	sit/stand	coughing	yawning	talking	putting_on_outerwear	vacuum	throwing	stretching	using_phone/tab	squat	reading_book
null	0.81	0.01	0.00	0.01	0.02	0.03	0.00	0.00	0.01	0.02	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.01	0.00	0.00	0.01
rinse_mouth	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
brushing	0.00	0.00	0.91	0.01	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
flossing	0.00	0.02	0.00	0.88	0.00	0.00	0.00	0.00	0.06	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
brush_hair	0.01	0.00	0.00	0.03	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.05	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
flush_toilet	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.09	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
open_door	0.01	0.00	0.00	0.00	0.00	0.01	0.86	0.00	0.00	0.00	0.03	0.00	0.00	0.02	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
washing_hands	0.00	0.08	0.00	0.00	0.00	0.05	0.00	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
eating	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
drinking	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.02	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pickup/putdown	0.02	0.00	0.00	0.00	0.03	0.00	0.01	0.00	0.01	0.00	0.80	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
pouring	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.79	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chopping	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.03	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
wiping_surface	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00
stirring	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.02	0.00	0.00	0.00
stationary	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.02
walking	0.01	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00
sit/stand	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.04	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
coughing	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
yawning	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.04	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.05	0.79	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00
talking	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.02	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.02
putting_on_outerwear	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.03	0.04	0.00	0.01	0.00
vacuum	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.05	0.00	0.00	0.00
throwing	0.03	0.00	0.04	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.79	0.03	0.00	0.00	0.00
stretching	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.86	0.00	0.04	0.00
using_phone/tab	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.07
squat	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.91	0.00
reading_book	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.88

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

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 on Mobile Platform

	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

We evaluated the ActSonic system's latency and overhead on the Google Pixel 7 mobile platform. Table 1 presents the inference time and various parameters of the ActSonic ResNet18 model evaluated on an in-the-wild 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.

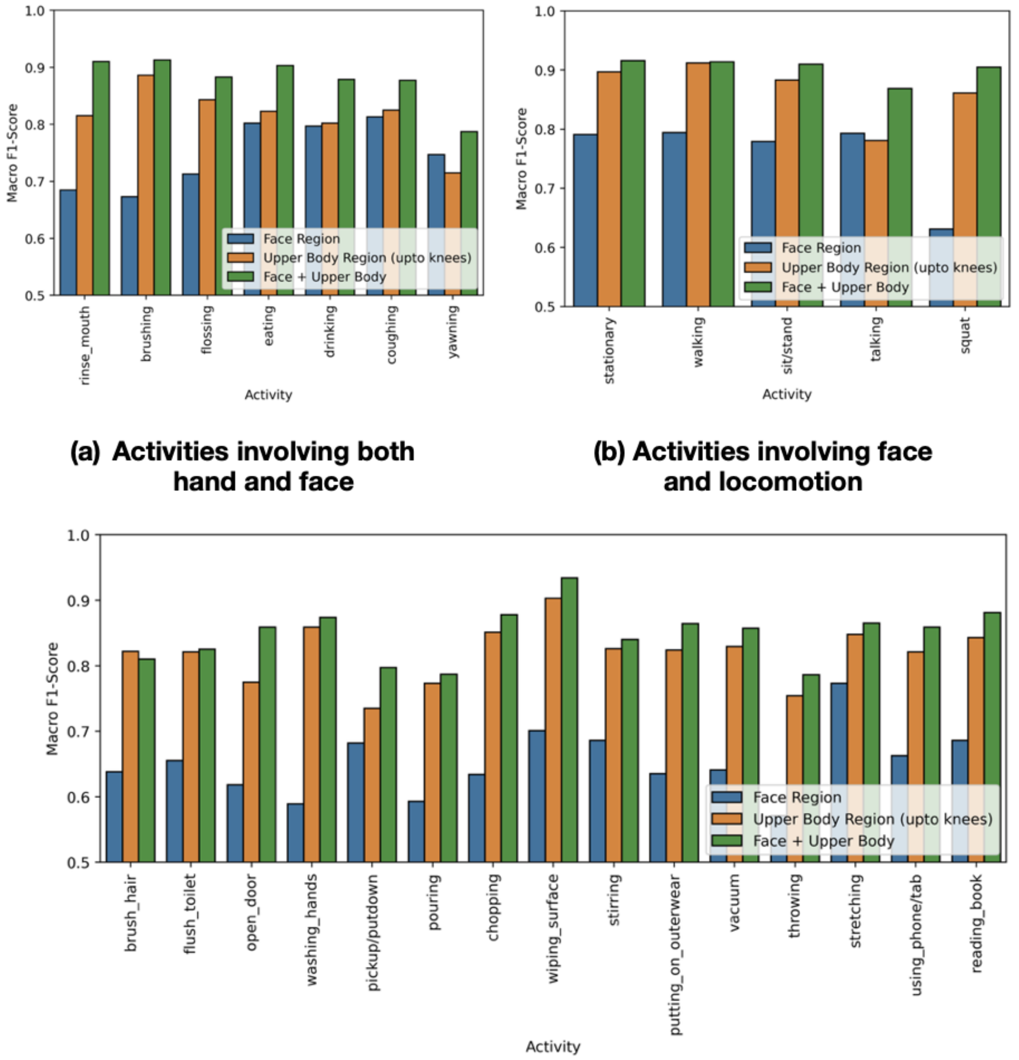
Furthermore, we utilized the Android Profiler [8] to evaluate mobile CPU usage and energy consumption. Table 1 indicates that while the quantized model exhibits lower accuracy compared to its non-quantized counterpart, it demonstrates lower system overhead. This assessment utilized a post-training quantization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

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.



(a) Activities involving both hand and face

(b) Activities involving face and locomotion

(c) Activities involving hands

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).

7.2 Performance Comparison of Different Deep Learning Encoders

We developed a self-supervised deep learning pipeline for ActSonic, utilizing ResNet18 [20] 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 [71] and the ConvLSTM architecture (ResNet18

Model Architecture	Number of Parameters (Approx.)	Prompted Sessions	In-the-wild 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).

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 in-the-wild 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.

7.3 Comparison of Performance with Prior Systems

We compare ActSonic with other activity recognition systems in Table 3. Many previous methods primarily focus on action recognition using passive data, such as passive acoustics, passive ultrasonic/infrasonic sensing, egocentric camera data, or inertial measurements from smartwatch data. However, these methods may not capture all actions comprehensively. Furthermore, most previous studies concentrate on smaller-scale experiments with specialized sensors or large-scale video-only benchmarks, with few considering in-the-wild settings.

ActSonic's utilization of an active acoustic sensing mechanism enables the capture of signals representing fine-grained movements that passive 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.

As shown in Table 3, most passive sensing approaches can detect audio events such as microwave usage, blender operation, or alarm clock sounds. However, these events do not necessarily imply that the user of the wearable system is engaged in specific activities. In contrast, ActSonic can recognize a wide variety of fine-grained body motion-based activities compared to other systems, achieving over 90% accuracy in naturalistic settings. Conversely, existing systems, as observed in Table 3, often demonstrate similar or worse performance, even with a smaller activity set or in controlled settings.

8 DISCUSSION

8.1 Saliency Analysis to Visualize Class Activation Feature Maps

Our user study demonstrates competitive accuracy in a wearable-based activity recognition system utilizing active acoustic sensing via echo profiles. Additionally, we conduct a saliency analysis to identify significant

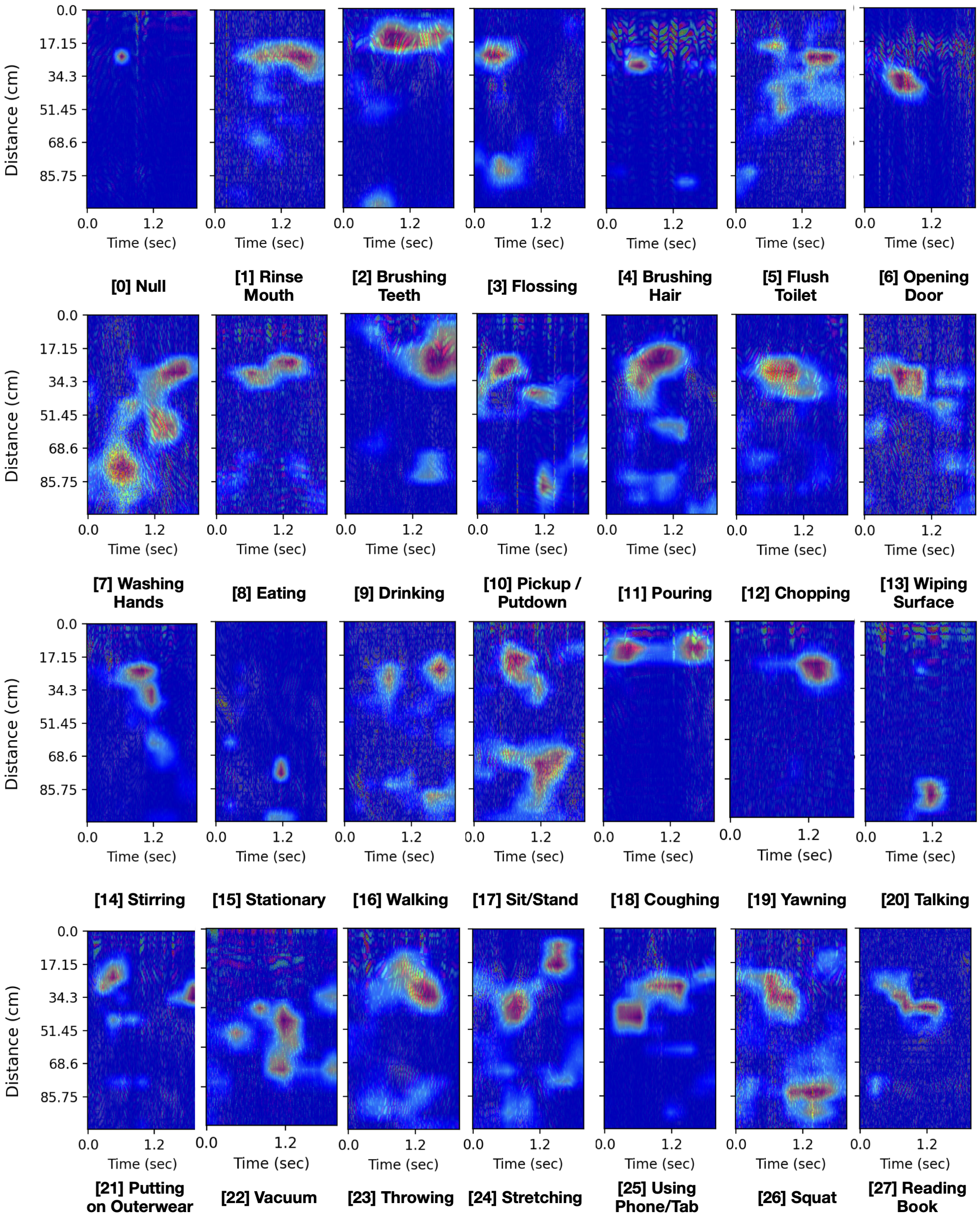


Fig. 11. Selected GradCAM heatmaps overlaid on differential echo profiles. As we have a 4-channel input and GradCAM aggregates heat maps by channel, we overlay the same (smoothed) heatmap across all four channels. Redder values from the smooth interpolation correspond to higher significance towards class prediction, while bluer values indicate lower significance.

System	Sensing Modality	Device Type	Number of Activities	Activity Examples	Performance (Accuracy)	Study Design
BodyScope [91]	Passive Acoustics	Bluetooth headset	12	eating, drinking, laughing, coughing	79.50%	Small-scale, In-the-wild, 4 activities
Ubicoustics [34]	Passive Acoustics	Commodity electronic devices with mic	30	chopping, baby crying, knocking, speech, alarm clock, etc.	89.60%	In-the-wild
PrivacyMic [23]	Passive Ultrasonic and Infrasonic sensing	Customized hardware board	10	mixer, microwave, kitchen sink, shredder, toilet, etc.	95%	Homes and commercial buildings
SAMoSA [58]	IMU and Subsampled Passive Audio	Smartwatch	26	drill, blender, microwave, coughing, toothbrushing, etc.	92.20%	At participants' home, activities performed according to instruction
DiffAct [50]	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. We also provide user study evaluation information in the table.

sections of the echo profile input for specific classes. Employing Gradient-weighted Class Activation Mapping (Grad-CAM) [73], we visualize the final convolutional layer of the ResNet18 encoder. Overlaying the feature heatmap on the 4 channels of acoustic flow or differential echo profiles confirms our model's capability to capture class-related motion within these channels. Figure 11 illustrates Grad-CAM for randomly picked sliding windows of all activities included in the ActSonic set.

In Figure 11, distinct activities exhibit activation in various regions of the echo profile sliding window. This observation validates that ActSonic's self-supervised model learns to focus on different explainable regions within the input sliding window to infer everyday 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 to infer the activity. Similarly, the activity of coughing illustrates hand movements in front of the face and the resulting motion pattern induced in the face due to coughing. Additionally, the activity of opening a door demonstrates hand movement to unlock the door, followed by movement to enter the room.

8.2 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 for echo profile calculation resulted in a slower frame rate than anticipated for real-time inference.

8.3 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 like HVAC, running water, TV, and ambient sounds, detailed in Sec. 6, ActSonic maintains consistent performance independent of environment settings. This resilience stems from its reliance on ultrasonic frequencies (18 KHz to 24.5 KHz), which surpass most environmental noise sources (recorded at frequencies below 7.5 KHz). For instance, noises like cafe chatter (63.8 dB), roadside traffic (69.0 dB), and loud music (71.5 dB) fell below ActSonic's operational range. Additionally, any overlapping frequencies in this range would result in significantly stronger signals near the eyeglasses, reinforcing ActSonic's robustness to environmental acoustic noise.

8.4 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.

8.5 Health Implication

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 [51]. The resultant intensity is 68 dB(A), well below the 85 dB limit set by NIOSH [61]. Research [59] 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. Future investigations will explore potential audibility among animals and children despite its inaudibility to adults for long-term usage.

8.6 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 :

8.6.1 Behavioral Journaling. We can utilize this system to journal various everyday activities for different purposes. For instance, one crucial step in combating eating disorders is to journal food intake behavior, often recorded manually. Previous eating journaling systems required multiple sensors, had low time resolution (e.g., recognizing a meal every 10 minutes) [76], or necessitated 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 the need for training data from a new user. This indicates that our system can be immediately deployed to facilitate eating journaling practices.

8.6.2 High-Resolution Behavior Data in the Wild for Health. ActSonic can track 27 everyday activities, many of which are related to health behaviors. Automatically logging these health-related activities can potentially provide researchers and clinical physicians an opportunity to better understand a user's activities in the wild for health purposes. For instance, the eating and drinking behavior can potentially be used to analyze the user's eating and drinking routines related to eating disorders and body hydration levels. Brushing teeth, flossing, and rinsing the mouth can be recognized with a very low error rate. This feature can be utilized by dentists to better track patient dental behavior.

8.6.3 Tracking other activities. In our study, we were only able to track 27 activities. However, our system has the potential to recognize other activities that involve body pose/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.

8.7 Limitations and Future Work

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

8.7.1 Scope of Activities 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, affecting system performance in real-world settings. Addressing this diversity might benefit from a larger dataset and a foundational deep learning feature extractor.

8.7.2 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.

8.7.3 Lack of Multi-Label/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 groundbreaking performance.

8.7.4 Lack of External Contextual Detection. Our focus on upper body motion doesn't consider the contextual environment. Integrating GPS or camera data could aid in understanding environmental affordances, and improving activity detection.

8.7.5 Lack of Access to Fine-Grained Hand or Head Movements. Our system currently captures upper-body motions only. Incorporating IMU sensors for head movements or additional acoustic sensors for precise hand tracking could enhance our approach through a multi-modal setup.

8.7.6 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 employing acoustic sensing on smart glasses. The extensive experiments involving 19 participants in real-world settings showcase

ActSonic's adeptness in distinguishing a diverse range of everyday actions across different environments. We envision ActSonic as a straightforward and efficient supplementary modality for egocentric action recognition, addressing concerns regarding privacy.

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] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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).

[17] 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.

[18] 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.

[19] 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

[20] 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.

[21] 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.

[22] 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

[23] 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.

[24] 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.

[25] 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

[26] 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

[27] 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.

[28] 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.

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

[30] 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.

[31] 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

[32] 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.

[33] 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.

[34] 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.

[35] 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

[36] 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

[37] 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.

[38] 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.

[39] 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).

[40] 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

[41] 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).

[42] 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).

[43] 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).

[44] 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

[45] 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.

[46] 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.

[47] 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

[48] 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

[49] 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.

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

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

[52] 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.

[53] 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.

[54] 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

[55] 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.*

[56] 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].

[57] 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

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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

[64] 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.

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

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

[67] 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.

[68] 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.

[69] 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

[70] 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.

[71] 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.

[72] 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.

[73] 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

[74] 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].

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

[76] 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

[77] 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.

[78] 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

[79] 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

[80] 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.

[81] 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.

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

[83] 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.

[84] 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.

[85] 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

[86] 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

[87] 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.

[88] 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.

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

[90] 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.

[91] 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.

[92] 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.

[93] 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* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3174011

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

[95] 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

[96] 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