SAMoSA: Sensing Activities with Motion and Subsampled Audio

VIMAL MOLLYN, Carnegie Mellon University, USA
KARAN AHUJA, Carnegie Mellon University, USA
DHRUV VERMA, University of Toronto, Canada
CHRIS HARRISON, Carnegie Mellon University, USA
MAYANK GOEL, Carnegie Mellon University, USA

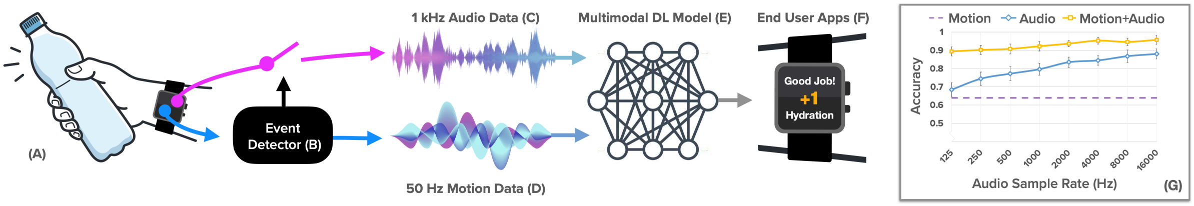


Fig. 1. SAMoSA senses a user's activity (A) using power-efficient and privacy-sensitive low-sample-rate (≤ 1 kHz) audio data (C) and 50 Hz IMU motion data (D). After a motion detector (B) triggers, these two sensor streams are passed to a multimodal deep learning model (E) to predict the wearer's activity, which can then be used to power end user applications (F). In this work, we study the privacy/power vs. accuracy tradeoff by testing different audio subsampling rates (G).

Despite advances in audio- and motion-based human activity recognition (HAR) systems, a practical, power-efficient, and privacy-sensitive activity recognition system has remained elusive. State-of-the-art activity recognition systems often require power-hungry and privacy-invasive audio data. This is especially challenging for resource-constrained wearables, such as smartwatches. To counter the need for an always-on audio-based activity classification system, we first make use of power and compute-optimized IMUs sampled at 50 Hz to act as a trigger for detecting activity events. Once detected, we use a multimodal deep learning model that augments the motion data with audio data captured on a smartwatch. We subsample this audio to rates ≤ 1 kHz, rendering spoken content unintelligible, while also reducing power consumption on mobile devices. Our multimodal deep learning model achieves a recognition accuracy of 92.2% across 26 daily activities in four indoor environments. Our findings show that subsampling audio from 16 kHz down to 1 kHz, in concert with motion data, does not result in a significant drop in inference accuracy. We also analyze the speech content intelligibility and power requirements of audio sampled at less than 1 kHz and demonstrate that our proposed approach can improve the practicality of human activity recognition systems.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *Interaction techniques*; Gestural input.

Additional Key Words and Phrases: Location-Aware/Contextual Computing, Sensors, Artifact or System

Authors' addresses: Vimal Mollyn, vmollyn@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Karan Ahuja, kahuja@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Dhruv Verma, dhruvverma@cs.toronto.edu, University of Toronto, Toronto, ON, Canada; Chris Harrison, chris.harrison@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Mayank Goel, mayank@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA.

2474-9567/2022/9-ART132
https://doi.org/10.1145/3550284

ACM Reference Format:
Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 132 (September 2022), 19 pages. https://doi.org/10.1145/3550284

1 INTRODUCTION

Human activity recognition (HAR) has numerous applications, including real-time task assistance, automated exercise tracking, rehabilitation, and personal informatics. Over the years, researchers have explored numerous modalities to sense a user's actions, and sound has proven to be a useful signal. Sounds resulting from physical activities, such as washing one's hands or brushing one's teeth, are often distinctive and enable accurate HAR. For example, Laput *et al.* built a model to classify sounds sampled at 16 kHz and achieved 89.6% average classification accuracy [34]. However, sampling audio at rates between 8 and 16 kHz carries a power consumption and compute cost. Moreover, these audio ranges capture human speech content and other sensitive information that users might not want to have recorded.

In a bid to protect user privacy, researchers have proposed to featurize recorded data [11, 41]. If done at the edge, featurization is considered privacy-sensitive, but it comes with considerable processing cost. Furthermore, an always-on acoustic activity recognition system increases the power burden; especially on resource-constrained devices such as smartwatches, where the battery cannot be made much larger. In response, we present *SAMoSA* - **S**ensing **A**ctivities with **Mo**tion and **S**ubsampled **A**udio. Our approach first uses power- and compute-optimized IMUs sampled at 50 Hz to act as a trigger for detecting the start of activities of interest. IMUs are ubiquitous, are heavily engineered for efficiency, and numerous prior works have shown the effectiveness of IMU data to detect user activities [6, 31, 35, 36, 60]. Once we detect the start of an activity using an IMU, we use a multimodal model that combines motion and audio data for classifying the activity (Figure 1). To further minimize the computation cost of processing audio data, we reduce the sampling rates to ≤ 1 kHz. Subsampling can be implemented directly and easily at the hardware level; *i.e.*, instead of post-processing fast-sampled audio data (*e.g.*, 44 kHz), the sensor is directly sampled at a reduced rate. This approach saves power needed to sample, move/store in memory, and featurize the data. Our approach was partially inspired by the always-on "Measure Sounds" (loudness) feature found on recent Apple Watch models. This software-based sensor relies on the microphone, but presumably senses at a very low rate so as to have minimal impact on battery life. At these lower rates, human speech is also unintelligible [5, 8, 15], thus offering a more privacy-sensitive approach.

We show in our evaluation that motion and sound signals are complementary and provide wide coverage of numerous activities of daily living. A similar approach (albeit without subsampled audio) is reportedly used in Apple Watch's handwashing recognition implementation, but there is no official documentation of the approach and its implementation. SAMoSA provides a generic, open-source approach that extends to 25 additional activities. Overall, this paper makes the following contributions:

(1) A practical activity recognition system that uses 50 Hz IMU along with audio sampled at ≤ 1 kHz, and yet still achieves performance comparable to models using more power-hungry and privacy-invasive 16 kHz audio data.
(2) A comprehensive suite of experiments, quantifying the efficacy of our system across 8 audio sampling rates, 4 contexts, 60 environments, and 26 activity classes. We also present a power consumption and a speech intelligibility study for different audio sampling rates.
(3) A new smartwatch sensor dataset with synchronized motion and sound data. We further showcase how our multimodal model resolves ambiguity among activities that are confused by a single modality alone.
(4) Open-sourced data, processing pipeline, and trained models to facilitate replication and further exploration and deployment in the field.

2 RELATED WORK

There has been extensive research done in human activity recognition (HAR) from sensors such as microphones [24, 34, 38], IMUs [6, 31, 35, 36, 60], cameras [62], powerline sensors [21], plumbing sensors [16], and various multimodal sensing approaches [1, 37, 40, 45, 49, 56, 59]. Please refer to [12, 20, 54, 57] for a detailed survey. In this section, we situate SAMoSA in the landscape of motion- and audio-based methods, as well as multimodal approaches for HAR.

2.1 IMU-Based Human Activity Recognition

A large number of activity recognition systems rely on inertial sensors such as accelerometers, gyroscopes, and magnetometers present in smartphones, smartwatches and other wearables to detect human activity. Over the years, motion-based activity recognition systems have graduated from detecting coarse human movements, such as walking, running and biking [31, 60] to detecting more fine-grained movements such as appliance usage [32, 36], subtle facial actions [61], and gestures [36]. These approaches include multi-device systems such as that of Shoaib *et al.* [53] that use the accelerometer from a smartphone in conjunction with one from a smartwatch to detect 13 activities, including smoking and drinking coffee. Such systems are lightweight and typically use simple statistical features computed from the input data streams. Single wearable approaches have also seen renewed interest with the advent of deep learning. Such approaches are computationally more expensive, but offer higher fidelity and better accuracy. For instance, by using a wrist-worn smartwatch accelerometer sampled at 4 kHz, Laput *et al.* [35] was able to detect fine-grained hand activities such as hand washing and scratching. Ashry *et al.* used an online Bi-LSTM deep learning architecture for continuous activity recognition on smartwatches [4]. We drew inspiration from these prior efforts in developing our IMU-based activity trigger and the IMU backbone of our multimodal activity classification model.

2.2 Privacy-Sensitive Audio-Based Human Activity Recognition

Apart from speech sensing, researchers have leveraged the ubiquity of microphones to build numerous audio-based HAR systems. Recently, with the advent of machine learning techniques that benefit from large existing datasets, there has been widespread success in building more general-purpose HAR systems [33, 34, 42, 55, 64]. However, all of these approaches use high sampling rates (≥ 16 kHz) and are therefore more computationally intensive and potentially more privacy invasive.

In response, researchers have explored the idea of privacy-sensitive activity recognition using audio. Larson *et al.* [38] proposed a cough sensing technique which only reconstructed cough sounds from the audio signal, while preventing speech from being reconstructed intelligibly. Chen *et al.* [11] proposed a vocalic selection and substitution method in order to reduce the intelligibility of speech while preserving environmental sounds. PrivacyMic [24] presented an approach using infrasonic and ultrasonic frequencies for activity recognition. Most related to our work is that of Liang *et al.* [41]. They presented a privacy-aware audio sensing framework across 15 classes for intentionally degrading audio frames by randomly replacing a subset of audio frames with nearby frames. Similar to their approach, we also degrade the audio intentionally through subsampling (*i.e.*, discarding data rather than replacing it), but safeguard against the loss of accuracy by using a multimodal learning framework (we also expand our activity set to 26 classes).

2.3 Multimodal Human Activity Recognition

Multimodal deep learning systems have seen widespread adoption in computer vision communities for tasks such as image captioning [3, 13, 25, 44], pose estimation [2, 51], and autonomous driving [9, 10]. Similarly, in recent years, human activity recognition has also seen an increased adoption of similar multimodal classification techniques [1, 19, 40, 45, 56, 59].

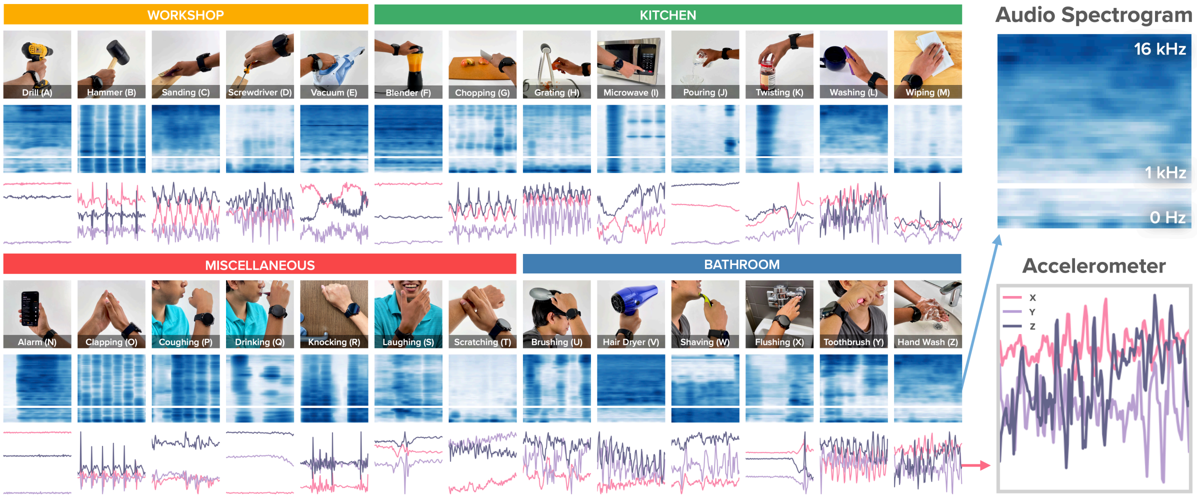


Fig. 2. SAMoSA is a human activity recognition system that uses IMU motion data sampled at 50 Hz and sound sampled at ≤ 1 kHz. The figure showcases the 26 different activities (with associated contexts) that our system can recognize, along with sound spectrograms (log scale, with 1 Khz cutoff denoted with a solid white line) and 3-Axis accelerometer streams.

Laput *et al.* [37] used a sensor board containing 9 sensors (accelerometer, magnetometer, light illumination, microphone, *etc.*) to create sensor-fusion powered "Synthetic Sensors" for a variety of context-sensitive applications. Radu *et al.* presented a case study of different challenges and approaches to multimodal deep learning for human activity recognition [49]. FingerSound [68] used sound and gyroscope data from a finger mounted ring to classify finger gestures. In Lukowizc *et al.* [43], the fusion of audio and IMU sensor data from multiple body worn sensors was used to detect a range of workshop activities. Most related to our present work are GestEar [7] and Ward *et al.* [63]. In GestEar, Becker *et al.* proposed the use of sound and motion from a smartwatch for gesture classification (*e.g.*, snapping, knocking, clapping). Similarly, Ward *et al.* proposed the use of multiple arm mounted microphones and accelerometers in order to classify assembly tasks [63]. In contrast to the latter systems, SAMoSA uses a commodity smartwatch, and thus constitutes a significant step forward in practicality and potential reach (*e.g.*, could be deployed with an over-the-air software update to even existing smartwatches). Moreover, our system detects an expanded set of events/activities compared to prior work.

3 DATASET

To initiate our investigations, we needed a labeled dataset with synchronized human motion and sound captured by a wrist-worn device. Such a multimodal dataset does not exist in the literature, and thus we collected our own.

3.1 Example Contexts and Activities

Rather than develop a new set of activities to study, we drew our example contexts and activities from highly related prior work [34, 35, 55]. From these papers, we selected activities that are general and usually produce motion or sound when performed. This meant that environmental sounds such as "baby crying" and "dog barking" found in some prior work (*e.g.*, [34]) were not included. Similar to [34], we also wished to balance our activities across different user contexts, and ultimately chose kitchen, bathroom, workshop, and miscellaneous. We binned our final set of 26 activities into these four contexts (Figure 2). As discussed later, we also recorded speech and transitional data between activities to act as a 27th *Other* class.

3.2 Data Capture

We collected data using a Fossil Gen 5 smartwatch running Google Android wearOS 2.23. We developed a custom wearOS app to collect and save/stream synchronized streams of the 9-axis IMU data (accelerometer, gyroscope and orientation) at 50 Hz and uncompressed audio at 16 kHz. *Post hoc*, we downsample the audio data into seven additional target rates (125, 250, 500, 1k, 2k, 4k, 8k Hz) for building our models and subsequent analyses. For this, we first subsample our 16 kHz audio signals to the desired target sampling rate by applying an n^{th} order subsampling scheme, wherein we simply keep every n^{th} sample and discard the rest. This closely mimics the subsampling scheme implemented in hardware.

3.3 Data Collection Procedure

We collected in-the-wild audio and inertial data from 20 participants (mean age 23.3, all right-handed) across 60 environments of varying use. We collected data in participants' homes with their appliances and tools. We only provided appliances in case they did not have a particular item (*e.g.*, not every participant had a handheld power drill). Participants performed activities belonging to a particular context in the location where they would most likely occur (*e.g.*, toothbrushing in the bathroom, chopping in the kitchen) to incorporate associated context background noise profiles (for example, ambient HVAC, water flowing through the pipes in the bathroom, *etc.*). Participants wore the smartwatch on their dominant arm. Each participant was asked to perform 26 activities across 4 contexts (Figure 2), with each activity repeated 3 times within each context. Participants were asked to perform these activities as they normally would and were given no further instructions. We encouraged them to use their own objects to perform the activities when applicable to increase variance in the data. In addition to the 26 activities, we also collected data for speech and when users were transitioning between activities. These instances acted as the *Other* class in order to test SAMoSA's motion-based event detector. In total, this procedure took about 90 minutes and each participant was compensated $20 USD for their time.

During data collection, we followed the best practices described in [30]. A trained experimenter oversaw the entire data collection and also acted as the activity annotator. First, a context was chosen at random, and then the trials of all the activities pertaining to that context were randomized. For example, in the workshop context, the 15 trials (5 activities × 3 repetitions per activity) were requested in a random order. This process was repeated until all the contexts were completed. For activities involving the use of an object (*e.g.*, hammer or hair dryer), each trial began with the participant picking up the object, performing the activity for a certain duration (ranging between 30 and 60 seconds) and then placing the object back down. Participants were not given explicit activity instructions on how to interact with the object and this led to useful variances in the data. For activities devoid of any objects (*e.g.*, laughing or clapping), the participants were simply asked to perform them. The experimenter annotated the start and end of each trial. All the data between the trials and the transition data was labeled as *Other*. In total, we have 14.2 hours of data across all participants. The dataset contains 5.9 hours of labeled activity data and 8.3 hours of in-transition *Other* data.

4 SYSTEM

The overall SAMoSA system architecture is illustrated in Figure 3. We start by first detecting activity events using only IMU motion data (Section 4.1). For this, we employ a lightweight Random Forest based model. Once the onset of an activity is detected, we trigger a multimodal deep learning model (Section 4.2). This model classifies the activity and further detects the end of the activity segment. Initially, the multimodal model is unaware of the activity context, meaning all 26 activity classes are applicable. However, if a sequence of activities that are

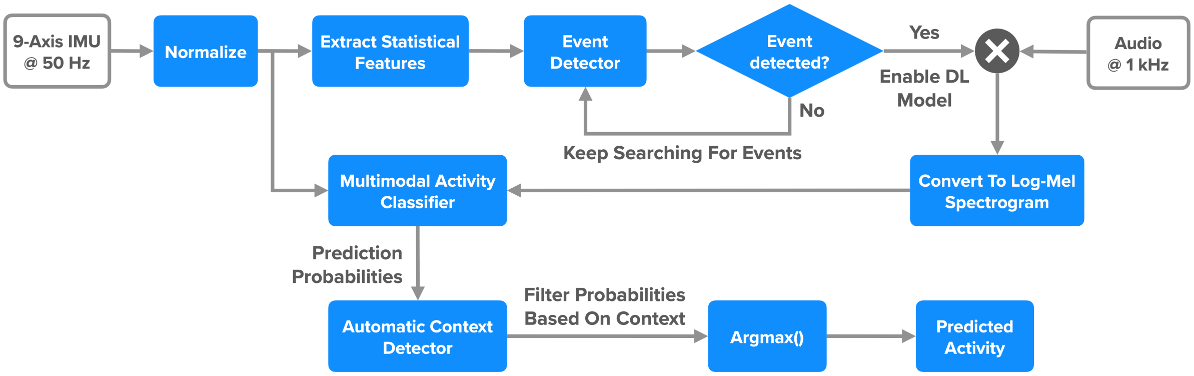

9-Axis IMU @ 50 Hz
Normalize
Extract Statistical Features
Event Detector
Event detected?
Yes
No
Keep Searching For Events
Enable DL Model
Audio @ 1 kHz
Convert To Log-Mel Spectrogram
Multimodal Activity Classifier
Prediction Probabilities
Automatic Context Detector
Filter Probabilities Based On Context
Argmax()
Predicted Activity

Fig. 3. Overview of SAMoSA's System Architecture.

inferred are consistent with a single context, our model will switch to a context-specific classifier (Section 4.3), which improves accuracy.

4.1 IMU-based Event Detector

4.1.1 Data Preprocessing. The first step of our pipeline is to preprocess our 9-axis IMU motion data (acceleration, rotational velocity, and orientation data across the x, y, and z axes). The smartwatch app logs orientation in a quaternion representation, which we convert to a rotation vector. Motion signals are then segmented into overlapping windows 2 seconds in duration with a stride of 200 ms. At 50 Hz sampling, this translates to an input array of 100 samples × 9 IMU values. Values are normalized between the ranges -1 and 1. This constitutes one motion instance, which are generated every 200 ms (*i.e.*, 5 Hz).

4.1.2 Featurization and Model. Prior works have explored several ways to extract features from IMU data, such as creating spectrograms [35], extracting statistical features [6, 36, 50], or extracting temporal features using a 1D CNN [14, 39, 66, 67]. In order to create a lightweight activity detection module, we employ a Random Forest Classifier trained on statistical features computed from the preprocessed IMU data stream. We extract 8 statistical features (mean, standard deviation, max, min, median, variance, skew, and kurtosis) for each of the 9 IMU values in each motion instance, resulting in a feature vector of size 72 (9 × 8). We choose these features due to their minimal computational cost for an always-on activity event detector. For the machine learning model, we use a Random Forest Classifier with 100 trees and a maximum tree depth of 10. The model was built using Scikit-Learn RandomForestClassifier [48], with a fixed random state and balanced class weights.

4.2 Multimodal Activity Classifier

4.2.1 Data Preprocessing. Our IMU data stream is already preprocessed as described in Section 4.1.1. Once the onset of an activity is detected, we preprocess the audio data. For this, we follow the method employed in [22]. We first run a short-time Fourier Transform (STFT) (window length 600 ms, stride of 30 ms) on the audio signal and then convert the resultant linear spectrogram into a 64-bin log-scaled Mel spectrogram. This always results in an audio instance of size 96 × 64 × 1 (approximately 3 seconds of audio data) regardless of the audio's sampling rate (which we vary from 125 Hz to 16 kHz, described in detail later). Also, regardless of audio sampling rate, we generate combined motion+audio data instances at 5 Hz (matching the output rate of our motion instances). We discard phase data, further obfuscating the audio content.

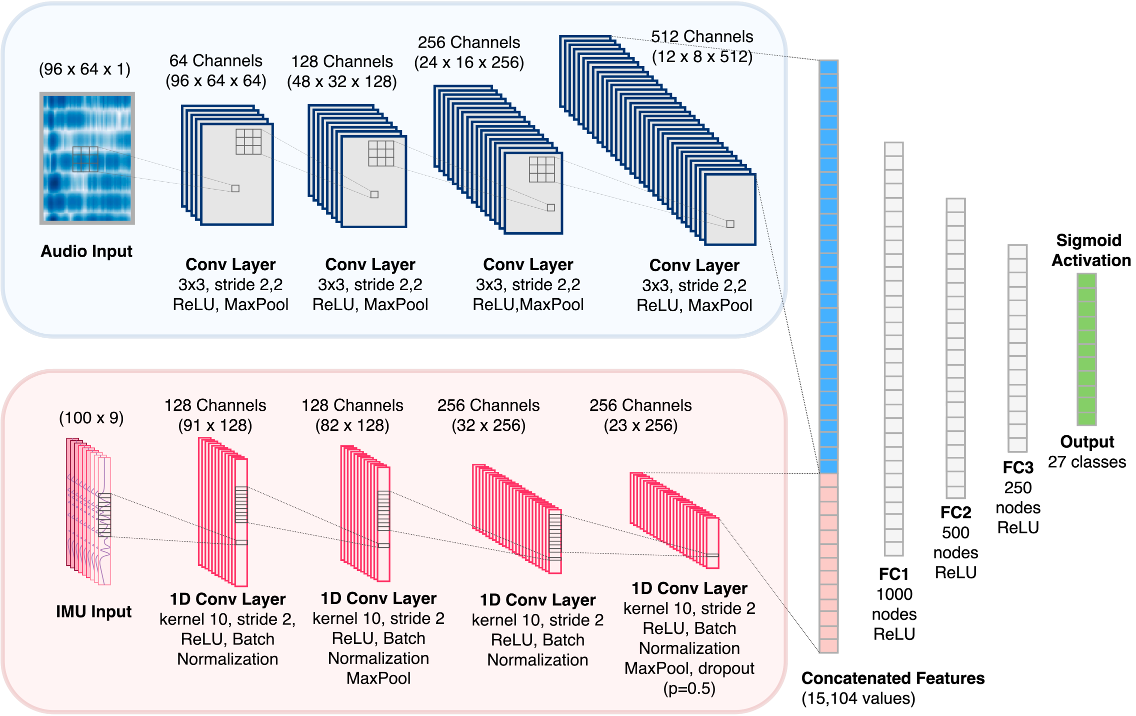


Fig. 4. Overview of SAMoSA's multimodal deep learning architecture. The model takes in a log-mel audio spectrogram along with a 9-DoF IMU motion instance as inputs to predict the activity. The sound modality is highlighted in blue and the motion modality is in red. The output (green) is the prediction probabilities for the 27 example classes.

4.2.2 Learning Architecture. Our multimodal deep learning model combines motion and audio data. To test the efficacy of each modality, we train three separate models: one for each data modality - motion, audio, and a third multimodal model that combines the two signals, which we now describe.

For the motion-only model, we employed a 1D CNN to learn temporal features from raw IMU data. A model overview can be seen in the IMU backbone (depicted in red) in Figure 4. The architecture consists of four 1D convolutional layers (kernel size = 10, stride = 1, depths = 128, 128, 256, 256, ReLU activation [46]) interspersed with batch normalization [23] and max pooling layers (pool size = 2), followed by a dropout layer (p = 0.5) to help with regularization. This creates a 2816 wide motion embedding which we feed into three fully connected dense layers and a final classification layer with sigmoid activations.

For the audio-only model, we build upon the VGG-16 architecture presented by Laput *et al.* [34]. The architecture (sound backbone depicted in blue in Figure 4) consists of four 2D convolutional layers (kernel size = 3×3, stride = 2, depths = 64, 128, 256, 512, ReLU activation) interspersed with max pooling layers. We replace the final fully connected classification layer with our own fully connected classification layer using sigmoid activations.

Finally, our multimodal model has two input backbones (Figure 4) resembling the feature extractors of each individual sensing modality, namely motion and audio. We chose to keep this similarity between models to enable a fair performance comparison between the different sensing modalities. The embeddings of each modality are fused together through a concatenation layer, before being fed into a set of fully connected dense layers (1000, 500 and 250 nodes, ReLU activation) and a dropout layer (p = 0.5). As with the other models, we add a final sigmoid activated classification layer.

4.2.3 Training Protocol. Our models were built using TensorFlow 2.3. We used the Adam [27] optimizer with a learning rate scheduler (ReduceLROnPlateau) with a starting learning rate of 0.001. Our audio model's backbone was warm-started with weights from the model released by Laput *et al.* [34]. As the motion data was kept consistent at 50 Hz and the audio input was binned into a log-mel spectogram, we used the same model architecture for different audio sampling rates. We checkpointed our models against training loss and stop training if the loss did not improve for 5 epochs. We found that given the small size of our dataset, checkpointing against training loss often resulted in the optimizer finding a more stable minima. Rather than training different models per-context, we train a single model across all our 27 classes (including the “other” class). For each context, we only look at the prediction confidences of classes pertaining to that context. A prediction is labeled as “other” if it belongs to the "other" class or has a confidence lower than the threshold for that context. This "other" class denotes the end of a given activity segment. Training was performed on an Nvidia Titan X GPU and took approximately 5 hrs (0.25 hrs per LOPO model) to train the multimodal CNNs.

4.3 Automatic Context Detector

In many cases, the context can be detected using other sensors on a smartwatch (*e.g.*, location of the device as inferred via WiFi or Bluetooth, or proximity to a fixed paired device like a smart speaker). However, this prior knowledge of the context may not always be available. In such cases, context could be inferred using the sound and motion data itself. We use the activities predicted by the SAMoSA model across time to detect the context. Specifically, we use a rolling window of 30 data instances, with each inference casting a vote for a specific context. The model selectively activates outputs for the winning context, typically improving stability and accuracy.

5 LIVE DEMO AND OPEN SOURCE

To create a proof-of-concept live demo, we streamed IMU and audio data captured by a smartwatch to a laptop (on the same WiFi network) for processing. Figure 5 showcases sample predictions of our multimodal model making use of 1 kHz audio and 50 Hz IMU, 16 kHz sound only model and 50 Hz IMU only model. Sample predictions were only displayed on the screen when prediction confidence was above a threshold (0.7). This system runs at roughly 24 frames per second on an Apple MacBook Air laptop with M1 processor. The inference time of SAMoSA's Random Forest IMU event detector and multimodal activity classifier is 4.8 ms and 41.2 ms, respectively. In the future, we envision a prototype similar to the hand washing detection feature on the Apple watch that runs locally. This can be achieved by engineering optimizations such as model compression and floating point quantization [17, 29].

Additionally, to enable other researchers to explore this domain, we have made our study data, architecture, trained models and visualization tools freely available at https://github.com/cmusmashlab/SAMoSA.

6 RESULTS

6.1 Event Detector Accuracy

We evaluated our event detection model in a leave-one-participant-out (LOPO) cross validation scheme. That is, in a given fold, we used data from 19 participants for training and tested on the remainder holdout participant. This was repeated 20 times (all combinations, results averaged). We treated the problem as a binary classification task, wherein all 26 labeled activity classes were binned together to signify the occurrence of an event and the *Other* class included segments from when users were transitioning between activities as specified in Section 3.3. Our Random Forest classifier made a classification (*Event* vs. *Other*) every 200 ms to detect the onset of an activity

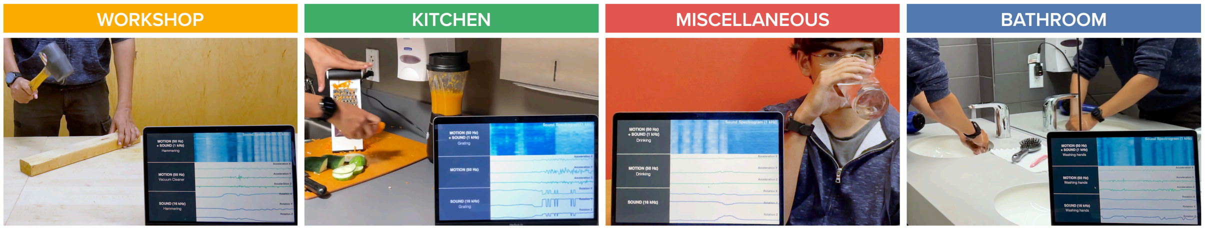


Fig. 5. Demo of SAMoSA running live in different contexts. Three models run concurrently for comparison, namely: the multimodal model (1 kHz audio + 50 Hz IMU), audio-only model (16 kHz audio) and the motion-only model (50 Hz IMU).

event. Our model was able to detect events with an average balanced F1 score of 0.88 (SD = 0.04). In terms of latency, we found an average onset latency (*i.e.*, the delay between the physical start of an event and when it is detected by the model) of 0.62 seconds, and an average offset latency (*i.e.*, delay between physical end of an activity and when the model detects the termination) of 0.16 seconds. We also computed these results across different machine learning models (refer to Appendix A.1).

6.2 Activity Recognition Accuracy

To test the efficacy of our activity recognition pipeline, we run it on our annotated activity segments. We designed our evaluation procedure in order to systematically isolate and analyze different factors. More specifically, we first analyze the effect of audio sampling rate on recognition accuracy. We then evaluate the efficacy of our multimodal model that combines subsampled audio and 50 Hz IMU data across different activities and contexts. Finally, we also evaluate the ability of our model to detect activity contexts. These are broken out into separate sections below.

All of our activity recognition models were evaluated in a leave-one-participant-out (LOPO) cross validation scheme. For accuracy metrics, we make use of segment level predictions, similar to *clip level* metrics reported by [34]. That is, for each segment pertaining to an activity, we return the top prediction based on the cumulative confidence of the model's output for the classes belonging to the given context. For frame-level metrics, refer to Appendix A.3. To improve generalizability, we do not train context dependant models, but rather train a single model across all activities with a sigmoid activated final layer. That is each output signifies whether an activity is occurring or not, thus modeling our classifier as a multi-label prediction rather than a multi-class classification. This enables us to selectively look at the model's outputs for activities present in a given context, without having to train context-dependant models.

6.2.1 *Analysis of Audio Sampling Rate on Accuracy.*

For each context, our accuracy across different audio sampling rates can be seen in Figure 6 (a mean of these plots can be found in Figure 1G). As expected, the recognition accuracy decreases in lockstep with a decrease in the audio sampling rate, especially in the region where speech is unintelligible (light green region). Overall, the sound-only model accuracy drops from 88.0% (SD = 8.6) at 16 kHz to 79.6% (SD = 10.9) at 1 kHz (Figure 6, blue lines). The performance keeps dropping as the sampling frequency goes below 1 kHz to 125 Hz. The drop in performance is the result of increased ambiguity between classes at lower sampling rates due to lower entropy. These results also translate to the full 27-class model; we observe the recognition accuracy decrease from 74.9% (SD = 6.9) at 16 kHz to 62.1% (SD = 8.2) at 1 kHz (Figure 6, yellow lines).

6.2.2 *Efficacy of Combined Motion and Audio Data.*

Although there is a similar decreasing trend in the performance of the motion+audio model, it is comparatively slower than the audio-only model (Figure 6). For

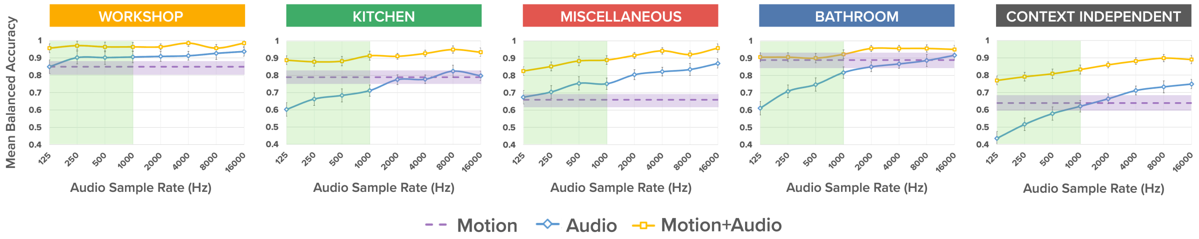


Fig. 6. Per-context mean classification accuracy of SAMoSA across different modalities and sound sampling rates. The IMU data is sampled at a constant 50 Hz for all models. Motion-only model performance does not vary with audio sampling rate as the models do not use audio data, and thus appear as flat lines (purple region illustrates standard error). The performance for audio-based models are roughly proportional to the sampling rate. However, when augmented with motion (motion+audio), the mean accuracy is consistently superior. Note that for improved readability, the y-axis starts at 0.4 for Mean Balanced Accuracy. Error bars shown here are standard error.

context-independent prediction, performance drops from 89.0% (SD = 5.9) at 16 kHz to 83.2% (SD = 8.5) at 1 kHz (Figure 6, far right plot). Moreover, when we average performance across all four contexts, the average performance of SAMoSA with 1 kHz audio and 50 Hz motion is 92.2% (SD = 7.9) (Figure 1G), and is marginally better than the average classification accuracy across contexts for 16 kHz audio-only model (88.0%). Thus, SAMoSA effectively combines motion information with low-sample-rate audio and outperforms a comparable model using 16 kHz audio. An interesting item to note from Figure 6 is that decreasing the sampling rate does not seem to have the same effect, in terms of performance, across all the contexts. The *Miscellaneous* context sees the biggest dip in accuracy when moving from 16 kHz to 1 kHz (−6.9%), while the *Kitchen* context sees the least drop (−2.0%). Recognition accuracies for activities such as *Pouring (J)*, *Scratching (T)*, and *Hand Washing (Z)* suffer the most, dropping by more than 20%, while the performances for *Drilling (A)*, *Sanding (C)*, and *Clapping (O)* remain stable even at 1 kHz sampling rates (see confusion matrices in Figure 7). Interestingly, even at a 125 Hz audio sampling rate, SAMoSA's multimodal model is comparable to 16 kHz audio model and consistently outperforms the baseline motion-only model across all contexts.

As the sampling rate of audio is reduced, the resolution of the information goes down, and more classes get confused with one another. This is evident in Figure 7. Sound spectrograms of activities such as *Hand Washing* and *Toothbrush* which were clearly distinguishable at 16 kHz, are much more similar below 1 kHz (Figure 8). In such cases, IMU data can provide valuable information to remove ambiguity (Figure 8, bottom row). However, it is important to note that motion alone cannot distinguish between different activities, as many of them have similar motion profiles (Figure 8; see *Microwave* and *Blender*). A multimodal approach can combine the "best of both worlds" and yield a more robust and generalizable classifier.

Inspecting the confusion matrices in Figure 7 offers some key insights. In the bathroom context, the 1 kHz audio-only model gets confused between classes that contain similar sounds. For example, *Toothbrushing (Y)* and *Washing Hands (Z)* get confused with each other about roughly 20% of the time, possibly due to the common "running water" background sound. On the other hand, the multimodal model is able to effectively disambiguate these two activities, even at a 1 kHz audio sampling rate, resulting in accuracies of 80.0% and 93.3%, respectively. Similarly, *Pouring (J)* and *Twisting (K)* have the lowest accuracies in the kitchen context of 47.5% and 35.6% respectively. The multimodal model boosts these to 86.2%, 76.2% respectively, which is sometimes even higher than the accuracies observed by the 16 kHz audio-only model. Finally, in the miscellaneous context, we observe that *Drinking (Q)* and *Scratching (T)* do not perform very well with the audio-only model (31.6% and 66.7% accuracies respectively) as they are predominantly silent activities, but with the multimodal model, they receive

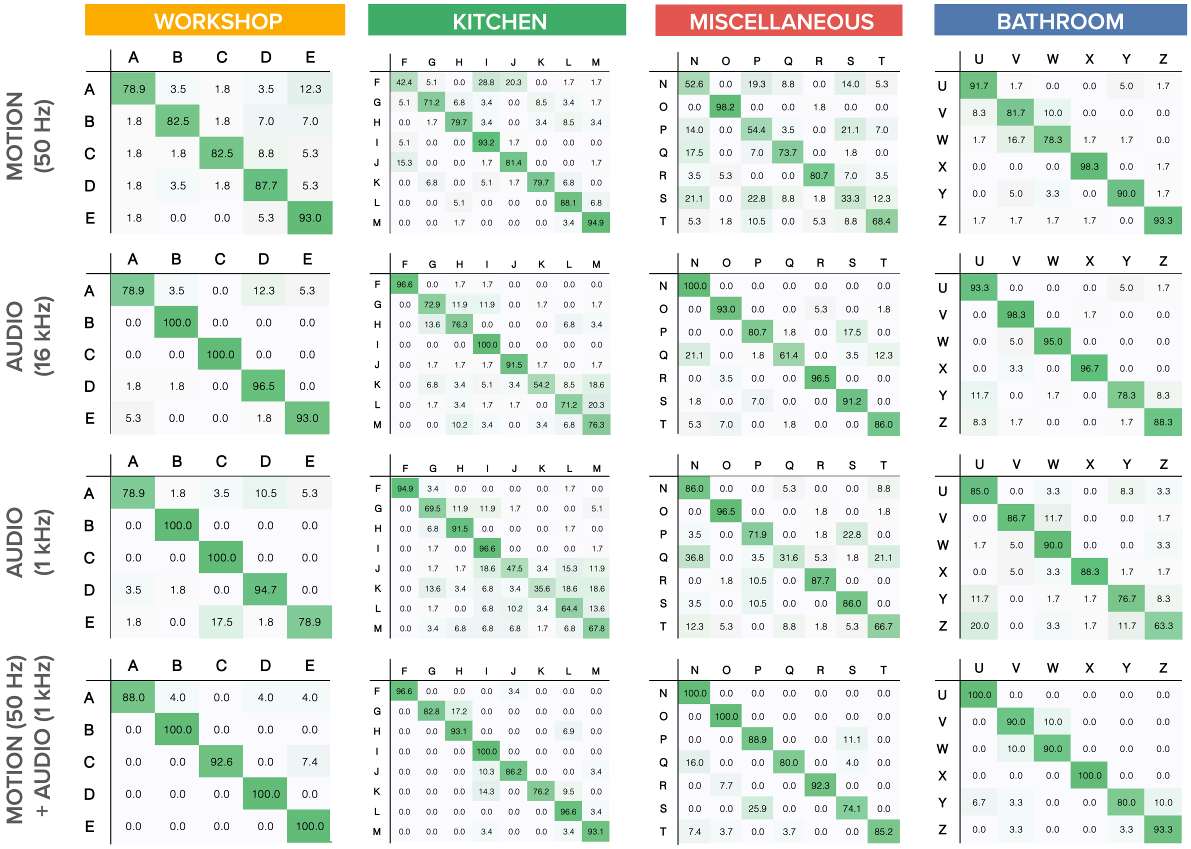


Fig. 7. Confusion matrices of SAMoSA across different contexts. The class legend can be found in Figure 2.

a significant boost in prediction accuracy (80.0% and 85.2% respectively). We also observed this using our live demo (Figure 5) wherein the sound model is not confident enough to make a prediction, but the multimodal one is sufficiently confident. Taken together, these results suggest a clear benefit from adding low-sampling-rate motion data to low-sampling-rate audio.

As noted previously, some activities have similar motion profiles. For instance, with our live demo (Figure 5, Workshop), we noticed that the motion model misclassified *Hammering (B)* as *Vacuuming (E)*. This is not surprising, as the two have very similar periodic motions (see also the confusion matrix 7 for motion). In another instance, the prediction probability of *Grating (H)* did not pass the threshold for our motion-only model and hence was not detected during the live demo (Figure 5, Kitchen). In both of these instances, the multimodal model helped in the disambiguation of these classes.

While a large number of activities benefit from including motion data along with audio, there are certain activities that are predominantly characterized by either audio or motion, which do not necessarily benefit from our multimodal model. For example, in the *Twisting (K)* activity, which is predominantly motion-based, we see that the motion-only model has an accuracy of 79.7% whereas the audio-only model has an accuracy of 54.2% at 16 kHz and 35.6% at 1 kHz. However, the combined multimodal model has an accuracy of 76.2%, which is only

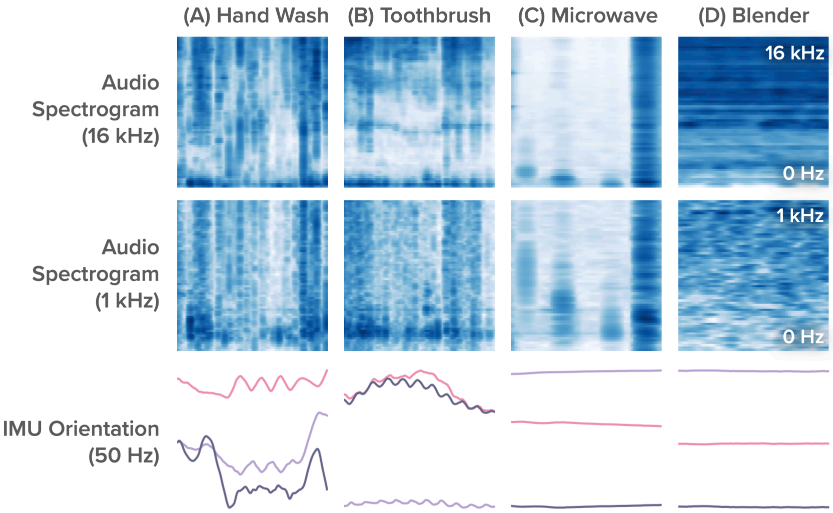


Fig. 8. Comparison of audio spectrogram and orientation signals across four exemplary activities. Activities with similar 1 kHz sound profiles (A and B) are distinguishable using motion. On the other hand, activities with similar motion profiles (C and D) can leverage 1 kHz sound to remove ambiguity.

about 3% lower than the motion only model. Similar, *Shaving (W)* has a high accuracy of 95.0% at 16k Hz sound, but degrades to 90.0% (similar to its accuracy at 1 kHz) in the multimodal model.

6.2.3 Automatic Context Detection. Our automatic context detection was also evaluated in a leave-one-participant-out (LOPO) cross validation scheme. For a particular holdout participant, we first select 10 random activity segments (corresponding to roughly 5 minutes of data in total) pertaining to a particular context. We randomly stitch together these segments to form a synthetic activity sequence. We use SAMoSA's multimodal classifier trained on 50 Hz IMU and 1 kHz audio data from the other 19 participants to classify each instance. We record the inferred activity's context and take the majority vote. We repeat this process 50 times per holdout participant with various permutations of contexts and activities. This results in a total of 1000 test combinations (20 participants × 50 random synthetic activity sequences). The automatic context detection accuracy (all combinations, results averaged) is 93.2% (SD = 8.3).

6.3 Full Stack Accuracy

As a final experiment – and one that better simulates real world performance – we ran our entire pipeline using the same LOPO procedure as above. Importantly, this includes the event detector as a "gate keeper" to model execution, as well as automatic context detection. We found an instance-wise accuracy of 69.7% (SD = 7.8). When run without automatic context detection, accuracy is 62.7% (SD = 7.9).

7 SUPPLEMENTAL STUDIES

We ran two supplemental studies to study the impact of audio sampling rate on microphone power consumption and on speech intelligibility (and by extension user privacy).

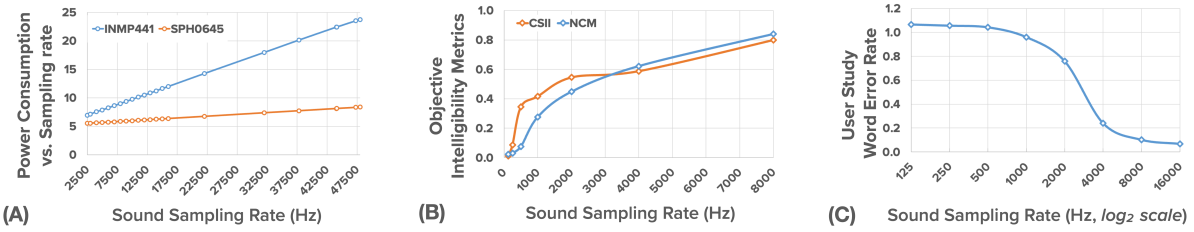


Fig. 9. Results of supplementary studies: (A) Power consumption (mW) *vs.* sound sampling rate of two digital MEMS microphones. (B) Two objective intelligibility metrics: Coherence and Speech Intelligibility Index (CSII) and Normalized-Covariance Measure (NCM) *vs.* sound sampling rate. (C) User study Word Error Rate (WER) *vs.* sound sampling rate - note the sharp increase in the word error rate below 1 kHz Sampling Rate.

7.1 Power Consumption vs. Audio Sampling Rate

A fundamental hypothesis underpinning our approach is that reducing the sampling rate of microphones on mobile hardware would result in power savings. Commercial-level optimization is hard to achieve in a research paper, but we can rely on some reliable proxies to see the effect of reduced sampling. As one point of comparison, we measured the power draw of two surface mount, digital MEMS microphones – Knowles SPH0645 [28] and InvenSense INMP441 [58] – equivalent to those found in smartphones and smartwatches. We measured the power draw of these microphones at varying sampling rates while transmitting data to an ESP32 board. We found a roughly linear relationship, illustrated in Figure 9A. More specifically, for the Knowles SPH0645 microphone, power consumption at 48 kHz was 8.37 mW, falling to 5.49 mW at 2.5 kHz. For the InvenSense INMP441 microphone, power draw fell from 23.74 mW to 6.91 mW as we lowered the sample rate from 48 kHz to 2.5 kHz. These power savings should compound as the data moves up the stack (*i.e.*, reduced memory and computational overhead due to less data).

7.2 Speech Intelligibility Study

In general, the most privacy-sensitive audio data is human speech content. Prior work has suggested a number of approaches to obfuscate audio data through various speech corruption schemes [11, 42]. We hypothesize that removing frequency bands encapsulating human speech should also help to preserve user privacy by rendering human speech unintelligible. While humans can hear audio frequencies ranging from roughly 20 Hz to 20 kHz, human speech occupies a relatively narrow frequency band from 400 Hz to 8 kHz [5, 8]. Thus, audio sampled at a rate below 800 Hz (the Nyquist rate for a 400 Hz audio signal) should theoretically render human speech unintelligible.

To more directly assess our hypothesis, we ran a study complementing prior seminal work [5, 8, 15]. We use objective intelligibility metrics drawn from the literature, and also recruited participants for a user study testing speech intelligibility. The latter serves as a real-world check that metrics cannot always provide. Objective speech intelligibility metrics such as Coherence and Speech Intelligibility Index (CSII) [26] and Normalized-Covariance Measure (NCM) [18] compute an intelligibility index by comparing a degraded signal with the original, noise-free signal. We compared the 96, 16 kHz sound samples with corresponding subsampled audio clips. The audio clips were selected from the LibriSpeech dataset [47], commonly used in benchmarking speech-to-text translation tasks. In Figure 9B, note the sharp decline in the intelligibility index (for both metrics) at around 1 kHz, as one would predict.

We further conducted an intelligibility study with 15 participants (self-reported native/fluent English speakers with mean age = 24.6 years) who were tasked with transcribing 24 audio clips (a subset of the 96 from the objective speech intelligibility study) from LibriSpeech. In this study, participants were presented with a tool that sequentially played 24 different human speech samples at 8 different sampling rates (3 clips for each sampling rate, ranging from 125 Hz to 16 kHz, in a randomized order). We evaluated the transcription quality using a popular speech-to-text translation metric: Word Error Rate (WER). The average WER across 15 participants and the eight audio sampling rates are plotted in Figure 9C. At 16 kHz we observe an WER of 6.8%, in line with prior works [52, 65]. However, as the audio sampling rate decreases, we observe a steep increase in WER. Notably, at a sampling rate of 1000 Hz, the word error rate is about 96%, which is in line with our hypothesis. Thus, once audio is subsampled at rates ≤ 1 kHz, speech data is largely unintelligible and certainly helps to preserve sensitive content.

8 LIMITATIONS AND FUTURE WORK

It is important to note that while our system has promising results, there are several limitations that will need to be addressed before SAMoSA is ready for deployment and consumer use. Foremost, the set of 26 activities is a small subset of the innumerable activities that occur in the real world. Future systems will either have to contend with larger class sets or focus on specific use cases (*e.g.*, hand washing, food intake, exercise) with robust false positive and equal error rate. As we found in our investigations, activities that sound the same can sometimes be separated by their motion data (and vice versa), which lends weight to multimodal approaches such as that of SAMoSA.

While SAMoSA offers automatic context recognition, the approach has several limitations. First, activities pertaining to the miscellaneous class can occur in any location. In such cases, rather than using rigid context class boundaries, different activity contexts would have to be merged dynamically to create new contexts on the fly. Furthermore, given the mobile nature of smartwatches, the context of the user could also change rapidly. In such cases, the context would need to be recomputed after a fixed interval of time, or when a change of location is detected through other sensors such as GPS, WiFi, or Bluetooth.

Since SAMoSA's event trigger is motion-based, it is inherently at a disadvantage while detecting sound-only events, such as an alarm clock. This often results in missed events, leading to degraded classification performance (See Section 6.3). However, this is a limitation that is necessitated by the power constraints of mobile devices, and could be rectified by designing highly-optimized, mobile-friendly activity classifiers, thus eliminating the need for a motion-based event trigger.

Overall, SAMoSA demonstrates competitive performance (accuracy of 92.2% across our four test contexts) when compared to prior work. However, the current results are still not sufficiently robust for immediate consumer use. Closing the gap from 92.2% to 99.9% (which is what consumers generally expect for interactive technologies) is a daunting challenge that will require future breakthroughs. However, there are many applications that do not need extremely high precision and accuracy. For example, step counting is not perfectly accurate, but users can still derive benefits from even approximate step counts. We believe that practical, deployable, and privacy-sensitive HAR systems with less than ideal accuracies will find similar uses to step counts (*e.g.*, mental health analysis, longitudinal behavior tracking).

We also note that privacy lies along a spectrum, and that even our low-sample-rate data could potentially reveal sensitive information, or in some cases, contain enough information for reconstruction. Our work is an important step in the direction of making HAR more privacy-sensitive, but it might not be enough for many users and applications. Thus work remains to further protect user privacy and adapt to different uses and users. A particularly promising idea we plan to explore in the future is to create sparse filter banks that permit classification identifying only a few characteristic and privacy-preserving frequency ranges per activity.

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] Karan Ahuja, Sven Mayer, Mayank Goel, and Chris Harrison. 2021. Pose-on-the-Go: Approximating User Pose with Smartphone Sensor Fusion and Inverse Kinematics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[4] Sara Ashry, Tetsuji Ogawa, and Walid Gomaa. 2020. CHARM-deep: Continuous human activity recognition model based on deep neural network using IMU sensors of smartwatch. *IEEE Sensors Journal* 20, 15 (2020), 8757–8770.

[5] Thomas Baer, Brian C. J. Moore, and Karolina Kluk. 2002. Effects of low pass filtering on the intelligibility of speech in noise for people with and without dead regions at high frequencies. *The Journal of the Acoustical Society of America* 112, 3 (2002), 1133–1144. https://doi.org/10.1121/1.1498853 arXiv:https://doi.org/10.1121/1.1498853

[6] Ling Bao and Stephen S. Intille. 2004. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive Computing*, Alois Ferscha and Friedemann Mattern (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–17.

[7] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: Combining Audio and Motion Sensing for Gesture Recognition on Smartwatches. In *Proceedings of the 23rd International Symposium on Wearable Computers* (London, United Kingdom) *(ISWC '19)*. Association for Computing Machinery, New York, NY, USA, 10–19. https://doi.org/10.1145/3341163.3347735

[8] Pranesh Bhargava and Deniz Başkent. 2012. Effects of low-pass filtering on intelligibility of periodically interrupted speech. *The Journal of the Acoustical Society of America* 131, 2 (2012), EL87–EL92. https://doi.org/10.1121/1.3670000

[9] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).

[10] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[11] 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* (Vancouver, British Columbia, Canada) *(MM '08)*. Association for Computing Machinery, New York, NY, USA, 733–736. https://doi.org/10.1145/1459359.1459472

[12] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Comput. Surv.* 54, 4, Article 77 (May 2021), 40 pages. https://doi.org/10.1145/3447744

[13] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[14] Stefan Duffner, Samuel Berlemont, Grégoire Lefebvre, and Christophe Garcia. 2014. 3D gesture classification with convolutional neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5432–5436.

[15] Grant Fairbanks and Frank Kodman. 1957. Word Intelligibility as a Function of Time Compression. *The Journal of the Acoustical Society of America* 29, 5 (1957), 636–641. https://doi.org/10.1121/1.1908992

[16] 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* (Orlando, Florida, USA) *(UbiComp '09)*. Association for Computing Machinery, New York, NY, USA, 235–244. https://doi.org/10.1145/1620545.1620581

[17] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. 2021. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630* (2021).

[18] Ray L Goldsworthy and Julie E Greenberg. 2004. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America* 116, 6 (2004), 3679–3689.

[19] Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. 2017. Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Inf. Fusion* 35 (2017), 68–80.

[20] Abhay Gupta, Kuldeep Gupta, Kshama Gupta, and Kapil Gupta. 2020. A Survey on Human Activity Recognition and Classification. In *2020 International Conference on Communication and Signal Processing (ICCSP)*. 0915–0919. https://doi.org/10.1109/ICCSP48568.2020.9182416

[21] 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* (Copenhagen, Denmark) *(UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 139–148. https://doi.org/10.1145/1864349.1864375

[22] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 131–135. https://doi.org/10.1109/ICASSP.2017.7952132

[23] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* (Lille, France) *(ICML'15)*. JMLR.org, 448–456.

[24] Yasha Iravantchi, Karan Ahuja, Mayank Goel, Chris Harrison, and Alanson Sample. 2021. *PrivacyMic: Utilizing Inaudible Frequencies for Privacy Preserving Daily Activity Recognition*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445169

[25] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[26] James M. Kates and Kathryn H. Arehart. 2005. Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America* 117, 4 (2005), 2224–2237. https://doi.org/10.1121/1.1862575 arXiv:https://doi.org/10.1121/1.1862575

[27] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[28] Knowles. 2022. *SPH0645 Digital MEMS Microphone*. https://www.digikey.com/en/products/detail/knowles/SPH0645LM4H-B/5332440

[29] Raghuraman Krishnamoorthi. 2018. Quantizing deep convolutional networks for efficient inference: A whitepaper. https://doi.org/10.48550/ARXIV.1806.08342

[30] Utkarsh Kunwar, Sheetal Borar, Moritz Berghofer, Julia Kylmälä, Ilhan Aslan, Luis A Leiva, and Antti Oulasvirta. 2022. Robust and Deployable Gesture Recognition for Smartwatches. In *27th International Conference on Intelligent User Interfaces*. 277–291.

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

[32] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[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* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 213–224. https://doi.org/10.1145/3242587.3242609

[35] Gierad Laput and Chris Harrison. 2019. *Sensing Fine-Grained Hand Activity with Smartwatches*. 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*. Association for Computing Machinery, New York, NY, USA, 3986–3999. https://doi.org/10.1145/3025453.3025773

[38] Eric C. Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N. Patel. 2011. Accurate and Privacy Preserving Cough Sensing Using a Low-Cost Microphone. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 375–384. https://doi.org/10.1145/2030112.2030163

[39] Song-Mi Lee, Sang Min Yoon, and Heeryon Cho. 2017. Human activity recognition from accelerometer data using Convolutional Neural Network. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. 131–134. https://doi.org/10.1109/BIGCOMP.2017.7881728

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

[42] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 17 (March 2019), 18 pages. https://doi.org/10.

1145/3314404

[43] 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 *International conference on pervasive computing*. Springer, 18–32.

[44] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090* (2014).

[45] 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

[46] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

[47] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[49] 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

[50] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. 2005. Activity Recognition from Accelerometer Data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3* (Pittsburgh, Pennsylvania) *(IAAI'05)*. AAAI Press, 1541–1546.

[51] Grégory Rogez, Maryam Khademi, JS Supančič III, Jose Maria Martinez Montiel, and Deva Ramanan. 2014. 3d hand pose detection in egocentric rgb-d images. In *European Conference on Computer Vision*. Springer, 356–371.

[52] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. 2017. English Conversational Telephone Speech Recognition by Humans and Machines. In *INTERSPEECH*.

[53] Muhammad Shoaib, Stephan Bosch, Hans Scholten, Paul J. M. Havinga, and Ozlem Durmaz Incel. 2015. Towards detection of bad habits by fusing smartphone and smartwatch sensors. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 591–596. https://doi.org/10.1109/PERCOMW.2015.7134104

[54] Wesllen Sousa Lima, Eduardo Souto, Khalil El-Khatib, Roozbeh Jalali, and Joao Gama. 2019. Human activity recognition using inertial sensors in a smartphone: An overview. *Sensors* 19, 14 (2019), 3213.

[55] Johannes A. Stork, Luciano Spinello, Jens Silva, and Kai O. Arras. 2012. Audio-based human activity recognition using Non-Markovian Ensemble Voting. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. 509–514. https://doi.org/10.1109/ROMAN.2012.6343802

[56] 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

[57] T Subetha and S Chitrakala. 2016. A survey on human activity recognition from videos. In *2016 international conference on information communication and embedded systems (ICICES)*. IEEE, 1–7.

[58] TDK-InvenSense. 2022. *INMP441 Digital MEMS Microphone*. https://invensense.tdk.com/products/digital/inmp441/

[59] 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

[60] Manuela M. Veloso and Subbarao Kambhampati (Eds.). 2005. *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*. AAAI Press / The MIT Press.

[61] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 129 (sep 2021), 28 pages. https://doi.org/10.1145/3478085

[62] Christian Vogler and Dimitris Metaxas. 1998. ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis. *Sixth International Conference on Computer Vision, 1998.* https://doi.org/10.1109/ICCV.1998.710744

[63] 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.

[64] 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.

[65] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2017. Achieving Human Parity in Conversational Speech Recognition. arXiv:1610.05256 [cs.CL]

[66] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-fourth international joint conference on artificial intelligence*.

[67] Mingzhi Zeng, Le T. Nguyen, Bo Yu, O. Mengshoel, Jiang Zhu, Pang Wu, and J. Zhang. 2014. Convolutional Neural Networks for human activity recognition using mobile sensors. *6th International Conference on Mobile Computing, Applications and Services* (2014), 197–205.

[68] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 120 (sep 2017), 19 pages. https://doi.org/10.1145/3130985

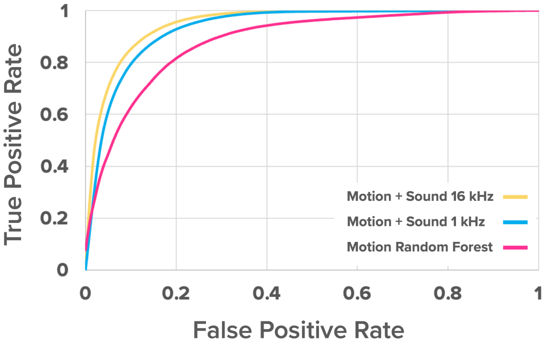
A SUPPLEMENTAL RESULTS

A.1 IMU-based Event Detection

Beyond the Random Forest classifier described in the main body of the paper, we also evaluated the performance of other popular statistical machine learning classifiers on our featurized IMU data. This included a Support Vector Classifier (RBF kernel) and a Logistic Regressor. The Support Vector Classifier and the Logistic Regressor had an average balanced F1 score of 0.85 and 0.82 respectively. This was comparable to the F1 score of our Random Foreset Classifier (0.88). The onset and offset detection times of the Support Vector Classifier were 0.59 and 0.19 seconds respectively, and that of Logistic Regressor were 0.61 and 0.30 seconds respectively. Our Random Forest Classifier has a comparable onset time of 0.62 seconds but a better offset time of 0.16 seconds.

A.2 Multimodal Event Detection

In the future, as edge neural net compute modules get further optimized in terms of performance and power, our multimodal activity recognition model could be executed continuously for event detection. For each classification, if a predicted activity is above a confidence threshold, we consider it an event. Else, if it's below a confidence threshold, or belongs to the *Other* class, we do not consider it an event. SoMoHAR's multimodal model is able to detect activity events with an Area Under the ROC (AUC) of 0.825 ($SD = 0.04$), within an average onset and offset time of 0.3 and 0.5 seconds respectively. Crucially, this is comparable to our sound-only model at 16 kHz which has an AUC of 0.820 ($SD = 0.03$), and an average onset and offset time of 0.2 and 0.3 seconds, respectively. For a detailed comparison between different modalities and audio sampling rates, please refer to Figure 10.

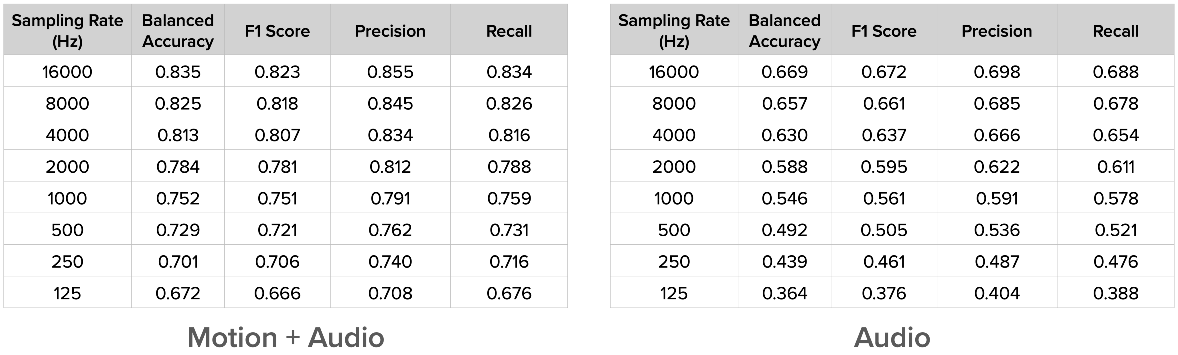


Modality	Sampling Rate (Hz)	EER	ROC AUC	F1 Score	Onset Delay (sec)	Offset Delay (sec)
Motion + Sound	125	0.423	0.874	0.793	0.217	0.554
Motion + Sound	250	0.471	0.900	0.819	0.210	0.616
Motion + Sound	500	0.505	0.938	0.862	0.241	0.497
Motion + Sound	1000	0.475	0.939	0.870	0.240	0.481
Motion + Sound	2000	0.469	0.949	0.882	0.153	0.439
Motion + Sound	4000	0.484	0.953	0.888	0.131	0.452
Motion + Sound	8000	0.503	0.951	0.887	0.116	0.467
Motion + Sound	16000	0.476	0.954	0.891	0.116	0.462
Sound	16000	0.487	0.940	0.869	0.094	0.379
Motion R. Forest	50	0.590	0.880	0.880	0.620	0.160

Fig. 10. (Left) ROC curves for detecting activity events using various models. (Right) Corresponding activity event detection metrics for models at different sampling rates.

A.3 Instance-Level Metrics for Activity Recognition

We also computed instance-level metrics for our activity recognition models across all activities (including the *Other* class). Figure 11 displays context-independent, instance-level performance in terms of balanced accuracy, F1 score, precision and recall. A similar trend to the segment-level metrics is observed, with both the audio-only and motion+audio models' performance decreasing in lockstep as the sampling rate is reduced. Furthermore, the multimodal model consistently outperforms those reliant on one modality alone.



Sampling Rate (Hz)	Balanced Accuracy	F1 Score	Precision	Recall
16000	0.835	0.823	0.855	0.834
8000	0.825	0.818	0.845	0.826
4000	0.813	0.807	0.834	0.816
2000	0.784	0.781	0.812	0.788
1000	0.752	0.751	0.791	0.759
500	0.729	0.721	0.762	0.731
250	0.701	0.706	0.740	0.716
125	0.672	0.666	0.708	0.676

Motion + Audio

Sampling Rate (Hz)	Balanced Accuracy	F1 Score	Precision	Recall
16000	0.669	0.672	0.698	0.688
8000	0.657	0.661	0.685	0.678
4000	0.630	0.637	0.666	0.654
2000	0.588	0.595	0.622	0.611
1000	0.546	0.561	0.591	0.578
500	0.492	0.505	0.536	0.521
250	0.439	0.461	0.487	0.476
125	0.364	0.376	0.404	0.388

Audio

Fig. 11. Instance-level metrics for activity recognition vs. audio sampling rate.