Detecting In-Person Conversations in Noisy Real-World Environments with Smartwatch Audio and Motion Sensing

ALICE ZHANG, CALLIHAN BERTLEY, DAWEI LIANG, and EDISON THOMAZ, The University of Texas at Austin, USA

Social interactions play a crucial role in shaping human behavior, relationships, and societies. It encompasses various forms of communication, such as verbal conversation, non-verbal gestures, facial expressions, and body language. In this work, we develop a novel computational approach to detect a foundational aspect of human social interactions, in-person verbal conversations, by leveraging audio and inertial data captured with a commodity smartwatch in acoustically-challenging scenarios. To evaluate our approach, we conducted a *lab* study with 11 participants and a *semi-naturalistic* study with 24 participants. We analyzed machine learning and deep learning models with 3 different fusion methods, showing the advantages of fusing audio and inertial data to consider not only verbal cues but also non-verbal gestures in conversations. Furthermore, we perform a comprehensive set of evaluations across activities and sampling rates to demonstrate the benefits of multimodal sensing in specific contexts. Overall, our framework achieved 82.0±3.0% macro F1-score when detecting conversations in the lab and 77.2±1.8% in the semi-naturalistic setting.

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

Additional Key Words and Phrases: Multimodal Classification, Audio Classification, Sound Sensing, Motion Sensing, Gesture Recognition, Non-verbal Communication, Wearable, Dataset, Smartwatch, Social Interactions, Human Activity Recognition

ACM Reference Format:
Alice Zhang, Callihan Bertley, Dawei Liang, and Edison Thomaz. 2025. Detecting In-Person Conversations in Noisy Real-World Environments with Smartwatch Audio and Motion Sensing. 1, 1 (March 2025), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Social interactions play a crucial role in shaping human behavior, relationships, and societies. It encompasses various forms of communication, such as verbal conversation, non-verbal gestures, facial expressions, and body language. Critically, lack of social interaction and loneliness are globally growing public health concerns, especially following the COVID-19 pandemic [39]. Prior research has shown that social isolation and loneliness are comparable to well-established risks for premature mortality, such as obesity and substance abuse [41]. Conventional methods of assessing an individual's social connections rely on retrospective clinician-rating surveys [46] or momentary self-report questionnaires called ecological momentary assessments (EMAs) [36]. Unfortunately, these methods have many shortcomings; they are susceptible to recall biases and often impose a significant burden on individuals. Moreover, these methods may not be appropriate for individuals who suffer from communication disorders, such as those with cognitive or language impairments. Consequently, a passive and universally-accessible method to sense and monitor social interactions

Authors' address: Alice Zhang, alice.zhang@austin.utexas.edu; Callihan Bertley, calbertley@utexas.edu; Dawei Liang, dawei.liang@utexas.edu; Edison Thomaz, ethomaz@utexas.edu, The University of Texas at Austin, Austin, Texas, USA.

ACM XXXX-XXXX/2025/3-ART
https://doi.org/XXXXXXX.XXXXXXX

in real-world settings would address these limitations and make a significant, positive impact in healthcare.

In this work, we advance a new computational direction in social interaction sensing that combines joint acoustic-inertial data to capture two key aspects of human communication: *in-person verbal conversations* and *non-verbal gestures*. Capturing non-verbal communication gestures is not only novel but highly relevant, since body movements in conversations can provide as much information as the spoken content itself [5, 33]. Furthermore, prior work focused on social interaction analysis has not considered this important dimension of social interactions [10, 29, 30, 34, 57, 61]. Additionally, we emphasize the analysis of social activities in acoustically challenging environments aligned with real-world settings.

To increase the practicality of our method, we leverage the rich and unobtrusive sensing capabilities of smartwatches. These devices are highly compelling for human-centered applications since they are generally comfortable to wear, socially acceptable, and increasingly ubiquitous. Importantly, they do not carry the stigma or burden of bulkier, customized sensors. The specific contributions of this work are:

(1) A novel multimodal approach to demonstrate the robustness and reliability of conversation detection in noisy and dynamic environments with acoustic noise levels ranging between 50-70dBA.
(2) An extensive evaluation of various modeling techniques and fusion methods showcasing the effectiveness of combining audio and inertial modalities for conversation detection, especially in environments with significant acoustic noise. We illustrate how multimodal approaches clarify situations that are confused by a single modality alone, with our framework achieving 82.0±3.0% and 77.2±1.8% macro F1-scores for detecting conversations in *lab* and *semi-naturalistic* studies respectively.
(3) A thorough set of analyses to illustrate the benefits of multimodal sensing across multiple contexts and audio sampling rates. To increase privacy protection in acoustic sensing, we show that the addition of inertial sensing to capture non-verbal conversational gestures can effectively supplement information lost in downsampled audio.
(4) A demonstration of the feasibility of our multimodal real-time sensing approach for smartwatches. We demonstrate that despite the model complexity, we can optimize and deploy the joint acoustic-inertial model onto a single commodity smartwatch with an average inference time less than 1s. We also perform a rigorous cost-benefit analysis of the smartwatch battery life and model performance across various sampling rates of our multimodal sensing method.
(5) A new, annotated dataset of social activities from 35 participants split into 11 groups across *lab* and *semi-naturalistic* settings with synchronized audio features and raw inertial data collected using an off-the-shelf smartwatch. This dataset enables research in disciplines ranging from understanding non-verbal communication in busy social contexts to improving technology interfaces. The data is available here.

2 RELATED WORK

2.1 Smartwatch Multimodal Sensing

The combined use of audio and inertial sensing with commodity smartwatches has been broadly explored in Human Activity Recognition (HAR) and Human-Computer Interaction (HCI) applications. Kim *et al.* used accelerometer and acoustic signals to classify 5 daily activities including eating, vacuuming, sleeping, showering, and watching TV [24]. Similarly, GestEar [6] jointly leveraged accelerometer, gyroscope, and acoustic signals to perform gesture classification on a limited set of simple gestures centered around snapping, knocking, and clapping.

Towards using audio and inertial sensing to recognize a greater number of activities, Siddiqui and Chan collected a dataset from inertial sensors and microphones as pressure-based sensors placed at the wrist to recognize a set of 13 daily life gestures and 1 relaxed gesture [50]. From this data, they hand-crafted and selected the most relevant features to gesture recognition using a mutual information-based algorithm that were then fed as inputs to classical machine learning models. Moreover, the ExtraSensory Dataset, collected from both smartwatches and smartphones of 60 participants, contains an even wider set of activities with user-labeled contexts [53]. The dataset contains IMU, location, phone state, and phone-recorded audio for classification of user contexts and activities, such as "at school" or "driving" [54].

Additional related work comes from Mollyn *et al.* who presented SAMoSa, a framework for recognition of 26 daily activities across several indoor environments using inertial signals and downsampled audio [35]. Likewise, Bhattacharya *et al.* collected synchronized audio and inertial data from a smartwatch for a set of 23 activities, such as writing and typing, for ADL recognition in *semi-naturalistic* and *in-the-wild* environments [7]. Liang *et al.* further demonstrated a teacher-student framework to build an IMU-based HAR model for greater accuracy in recognizing this set of activities for ADL recognition [28]. During training, the IMU model is augmented with acoustic knowledge, and once trained, the model only uses motion inputs for inference.

Contrasting against this previous body of research on multimodal sensing through a smartwatch, our work focuses exclusively on the detection of in-person conversations in challenging real-world scenarios. We leverage inertial data to capture non-verbal behaviors from in-person conversations to aid in conversation detection when the audio modality alone is otherwise insufficient for the task due to background sounds. Another difference is that in prior work, participants were instructed to wear the smartwatch on a specific hand (usually dominant hand) to better capture hand-based motion patterns of activities [3, 6, 24, 28, 35]. In our studies, however, participants were free to choose the hand on which to wear the smartwatch.

2.2 Social Interaction Sensing

Existing methods for sensing face-to-face social interactions can be clustered into two categories: smartphone-based methods and wearable sensor-based methods.

2.2.1 Smartphone-Based Methods. Smartphone-based methods utilize sensors in smartphones to detect and analyze social interactions. SocialWeaver [30] and DopEnc [61] use Bluetooth and doppler profiling, respectively, with signals transmitted and received between individuals' smartphones to determine the proximity between two individuals and infer whether they are engaged in conversation. Similarly, Crowd++ [59] uses audio collected from smartphones to estimate the number of speakers in a group. However, these smartphone-based methods do not work as intended when individuals do not carry their smartphones on body but instead in a purse or backpack, as is commonly observed in some populations when not actively using the phone [45]. Therefore, the assumption of these methods that smartphones are primarily carried on-body limits these methods' scope of application.

There are also collaborative sensing methods involving multiple smartphones for social interaction analysis, such as SocioPhone [25] and Darwin [34], which use individuals' voice signals captured by microphones across multiple smartphones. Data from multiple devices are shared to detect conversational turns as in SocioPhone or perform speaker recognition as in Darwin. Additionally, Liu *et al.* leverages a smartphone in coordination with multiple on-body inertial sensors to monitor in-person interactions [27]. However, employing multiple devices can restrict their ease of scalability.

2.2.2 Wearable Sensor-Based Methods. An alternative to smartphone sensing is using custom or commercial wearable sensors to infer interactions among individuals. Previous studies have developed custom hardware devices, such as sociometric badges with infrared sensors (IR) [38] or active radio frequency identification (RFID) tags [10], worn around the neck with a lanyard. Both the IR sensor and RFID tag identify instances when two individuals are directly facing each other within close proximity (<1 meter), indicating that the individuals are engaged in an interaction. However, these wearable badge-based methods require each individual in the interaction to have and to wear their own badge. Hence, these methods are not easily scalable to social interaction sensing in general populations.

Rahman *et al.* [44] and Bari *et al.* [3] both developed conversation detection methods from respiration signals collected by a chest band worn around a speaker's chest. Bari *et al.* further leveraged electrocardiogram sensors within the chest band and inertial sensors on a wristband worn on the dominant hand to specifically detect stressful conversations. However, continuously wearing a chest band for everyday sensing can be inconvenient and interfere with daily activities.

Most recently, commodity devices have been used to recognize face-to-face social conversations. Commercially, Apple AirPods Pro feature a Conversation Awareness mode that automatically reduces media volume and background noise upon detecting the user's speech [2]. Off-the-shelf smartwatches have also been used to detect and quantify face-to-face social interactions. In studies by Liang *et al.* [29] and White *et a*l. [57], the microphone of a commodity smartwatch captures acoustic features that are used to detect instances of in-person conversations. Additionally, White *et al.* specifically requires collecting voice samples and developing a voiceprint unique to each user during model training in order to compare the input audio to the pool of known speaker identities during model inference.

Our work differs from these prior works in that it leverages multiple sensors within a single commodity smartwatch to infer instances of face-to-face conversations. Unlike White *et al.* which requires customizing the model to specific users, our framework is speaker agnostic. This reduces the required pre-training overhead and increases user privacy as we do not need to collect voice samples for speaker identification or verification - voice samples that if misused or leaked could be leveraged for voice spoofing. The current work also seeks to improve model performance in especially loud and busy environments, such as restaurants or bars, where background conversations can be confused for a device user's conversations.

2.3 Speech Processing Tools

Speech processing tools have recently grown significantly in their capabilities. Speech processing tasks include speaker diarization, speaker recognition and verification, speech recognition, and more. Whisper, for instance, is an automatic speech recognition (ASR) system that enables real-time transcription and translation in multiple languages [16, 43]. *pyannote.audio* is a speaker diarization pipeline that recognizes who spoke when in a given segment of audio [8]. Meanwhile, intermediate embeddings, such as i-vectors and x-vectors, extracted from deep neural networks allow for recognizing speakers [47].

While these tools provide state-of-the-art performance for their respective speech tasks, these tools have limited applicability to analyzing social interactions. Furthermore, non-verbal communication, such as facial expressions, body language, gestures, and posture, play a significant role in face-to-face communication and are analyses beyond what current speech processing tools can provide. Through the addition of inertial data, our work is different than existing speech processing works as we aim to investigate the dynamic and interactive processes involved in social interactions, namely gestures and body movements. Lastly, we do not employ any natural language processing on the input audio, which increases the privacy of the user's spoken content.

3 CONVERSATION MODELING

In this work, we aim to sense social interactions, specifically face-to-face conversations. Previous works in detecting face-to-face conversations have varied in their approach to modeling conversations as there is significant variability in real-world social interactions [55]. Rahman *et al.* [44] defined conversation episodes to consist of user speaking and listening events while other works used conversational turn-taking as the fundamental unit of conversations [23, 25, 29]. Not all speech constitutes conversation; however, there is common agreement in literature that conversations across languages and cultures are characterized by turn-taking [14, 22, 48]. Thus, we also formulate this task around conversational turn-taking involving the device user. We approach the task as a three-class classification problem in line with existing literature and previous work [29]. The three classes are: 1) conversation, 2) other speech, and 3) background noise.

The *conversation* class is defined to be instances where spoken communication with turn-changes occurs between the participant wearing the smartwatch and at least one other participant in the study. The *other speech* class is defined as instances where spoken communication does not involve the participant wearing the smartwatch or foreground speech by the smartwatch user that does not contain turn-changes. Additionally, this class captures instances of when a participant wearing a smartwatch stops participating in a group conversation. Lastly, the *background noise* class contains instances where there is no face-to-face spoken communication in the foreground. That is, this class captures speech from music, TV, or spoken communications in the background.

4 DATA COLLECTION

To develop and evaluate our approach, we collected a labeled dataset with synchronized audio and inertial data from a wrist-worn device during social activities in acoustically challenging environments. This multimodal dataset does not exist in literature, which prompted us to create one. This dataset can further open research avenues in understanding movement patterns during face-to-face communication, improving human-machine interaction by recognizing social cues, and more. This section presents the data collection process realized through two IRB-approved user studies - one performed in the *laboratory* and one performed in *semi-naturalistic* settings. In each study session, groups of two to four participants engaged in a set of social activities. We first present the hardware setup, followed by the data collection and annotation protocols.

4.1 Hardware Setup

We used one Fossil Gen 4 smartwatch and one Fossil Gen 5 smartwatch to collect data from participants. Both smartwatches are equipped with a Qualcomm Snapdragon 3100 processor and driven by Google's WearOS operating system. The smartwatches also have built-in accelerometer, gyroscope, and microphone sensors. On both watches, we collected audio and inertial data synchronously and saved data locally on the device using a custom-developed Android application. Lossless audio data was recorded at a sampling rate of 16kHz and 6-axis IMU (accelerometer and gyroscope) data was recorded at a sampling rate of 55Hz. We verified that differences in data collected between the two smartwatches were negligible. Post hoc, we downsampled the audio data into two additional sampling rates (1kHz and 2kHz) for model development and analysis, as further described in section 5.1 and used in section 7.3. Researchers also recorded a video of each study session in its entirety for reference during the annotation process.

Table 1. Study participant and group details (L: Left, R: Right, M: Male, F: Female, SW: Smartwatch).

Group #	Group Size	Setting	SW User 1			SW User 2		
			Handed-ness	Watch Hand	Gender	Handed-ness	Watch Hand	Gender
1	3	Lab	R	L	M	R	L	M
2	3	Lab	R	L	F	R	L	M
3	2	Lab	R	L	M	R	L	M
4	3	Lab	R	L	M	R	L	M
5	3	SN (lobby)	R	L	M	R	L	M
6	3	SN (lobby)	R	L	M	R	L	M
7	3	SN (lobby)	L	L	F	L	L	F
8	4	SN (outdoors)	R	L	F	L	R	F
9	3	SN (lobby)	R	R	M	R	L	F
10	4	SN (lobby)	R	R	F	R	L	M
11	4	SN (outdoors)	L	L	F	R	L	M

Table 2. Distribution of participants' handedness and watch wrist.

		Watch Wrist	
		Left	Right
Handedness	Left-hand Dominant	3	1
	Right-hand dominant	16	2

4.2 Data Collection Protocol

We collected data from 11 groups of participants across *lab* and *semi-naturalistic* settings for a total of 35 participants. Each group had a unique set of participants that did not overlap with any other groups. The *lab* setting was a quiet, acoustically-controlled environment. The *semi-naturalistic* settings included the lobby of a busy academic building in which there were background conversations and non-speech sounds from sources such as elevators and rolling utility carts, and an outdoors patio café in which there were background conversations and non-speech sounds from sources such as wildlife and engines.

To better quantify the acoustic characteristics of the data collection environments, we measured the A-weighted, equivalent continuous sound level (LAeq) of the environments without participant activity using the National Institute for Occupational Safety and Health (NIOSH) Sound Level Meter application on an iPhone, which is compliant with sound level meter standards [11, 51]. LAeq, measured in decibels, is the average sound energy over a period of time that emphasizes frequencies perceived by humans and is commonly used as a standard metric of noise levels. The *lab* setting without participant activity had an average LAeq of 50.7dBA. The lobby of the *semi-naturalistic* setting had an average LAeq of 70.2dBA and the outdoors patio café had an average LAeq of 60.3dBA. For reference, rainfall is around 50dBA, a normal conversation is around 60dBA, and a washing machine is around 70dBA [13]. These sound levels show the acoustic variations of the data collection environments and specifically highlight the acoustically challenging nature of the *semi-naturalistic* environments.

Each group consisted of two to four participants. Participants were diverse in gender and cultural representation, with participants of American, Chinese, Indian, and Australian backgrounds. This is important as communication styles, especially non-verbal communication, vary across cultures [32]. etails on the composition of the groups are in Table 1 and Table 2.

Within each group, two participants wore the data-collecting smartwatches. To increase the ecological validity of the study, the two participants were instructed to wear the smartwatch on whichever wrist they would normally wear a watch. All group participants clapped at the beginning of the recording process to synchronize audio and inertial data for data annotation and processing. Within their groups, a researcher asked all participants to perform the following activities:

(1) **Group Conversation:** All participants of the group played a NASA decision-making survival game while sitting [19]. In the *lab* sessions, participants sat in chairs around a whiteboard, while in semi-naturalistic sessions, participants sat in chairs around a table. A participant not wearing the smartwatch was tasked with recording the group's responses on a whiteboard in the *lab* sessions and on a sheet of paper in the *semi-naturalistic* sessions. To emulate instances where individuals are situated in group conversations but not actively participating in the discussion, one participant wearing a smartwatch was instructed to discontinue speaking partway through the game and only listen.
(2) **Group Conversation While Eating:** All participants were given snacks and instructed to chat with each other on any topic of their choosing while eating their snacks. Many social gatherings involve food and take place at restaurants or other venues that can be acoustically noisy. The purpose of collecting data with this activity was: 1) to simulate these louder environments in which social interactions occur and 2) to capture hand movements from eating in order to understand differences in hand movements due to speech-related gestures versus eating. Similar to activity 1, one participant wearing a smartwatch was instructed to discontinue speaking partway through the group conversation and only listen while eating.
(3) **Listening to Music:** Using a speaker, the researcher played music from a variety of musical genres. All participants listened to the music for 2-3 minutes.
(4) **Reading Out Loud:** Each participant wearing a smartwatch read out loud a random passage from one of three non-fiction books for two minutes. Then, a researcher played music different than the music played in activity 3, while the participant continued reading for another 2 minutes.
(5) **Watching TV:** All participants watched two 3-5 minute video clips from a set of pre-selected clips from TV shows, talk shows, sportscasts, and documentaries, all of which contained conversations or narrations. Participants set the volume of the video clip playback, and conversation between participants was allowed.

This set of activities was chosen for being acoustically challenging yet representative of activities that take place in daily life. In total, we had a total of 35 participants across all study sessions and collected audio and inertial data from 22 participants. Our annotated dataset contains a total of 14.6 hours of audio and inertial data. Figure 1 shows screenshots of videos recorded during each study session, highlighting the setting and nature of activities.

4.3 Data Annotation Process

After data collection, one researcher (one of the paper authors) manually annotated participants' audio and inertial data. We used ELAN [58] to annotate audio while referencing the recorded video for ground truth and followed an annotation scheme established in prior works as discussed in section 3. We initially examined the social event detection problem at a granularity of 10-second segments. Consequently, we assigned a label, either *conversation*, *other speech*, or *background noise*, to each 10-second segment of audio and inertial data. For 10-second segments that contained more than one class of activity, we assigned the segment with the class that occupies the majority of the

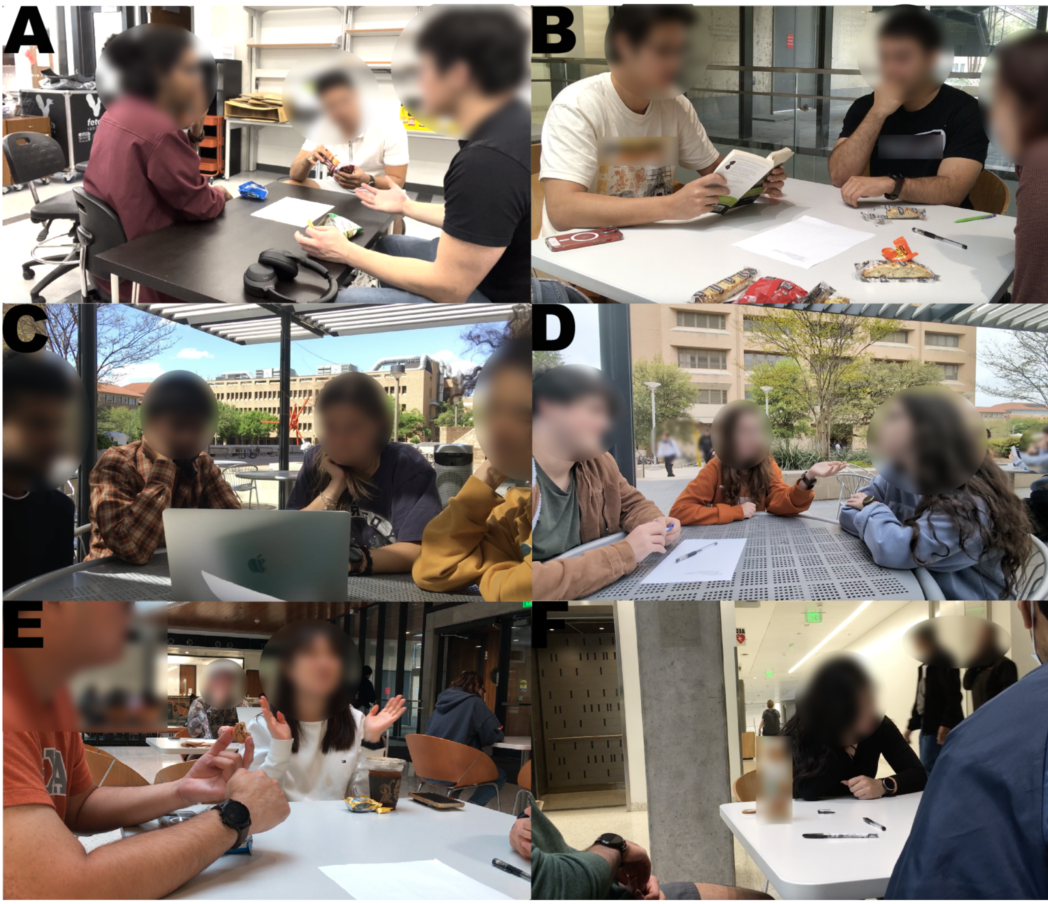

Fig. 1. Participants from different sessions performing group activities performed in the *lab* and in a *semi-naturalistic* setting. A: Participants having conversations while eating in a *lab* setting. B: Participants reading out loud in a *semi-naturalistic* setting. C: Participants watching a video in an outdoors *semi-naturalistic* setting. D: Participants playing a team building exercise in a outdoors *semi-naturalistic* setting. E: Participants having conversations while eating in an indoors *semi-naturalistic* setting (two participants not shown). F: Participants playing a team building exercise in an indoors *semi-naturalistic* setting (two participants not shown).

10-second segment. For instance, in a 10-second segment that contains 7 seconds of conversation and 3 seconds of audio from a video clip, the 10 second segment is assigned the *conversation* label.

As will be discussed in detail in section 7.1.1, we found that our approach achieves optimal performance at window lengths of 30 seconds through a sensitivity analysis. Therefore, we aggregated our 10-second labels and applied them to 30-second segments as appropriate following the same method of assigning the 30-second segment with the label of the majority-duration class. The remainder of our paper discusses our framework with 30-second window lengths.

5 SOCIAL EVENT DETECTION

In this section, we present the data preprocessing and models developed for our conversation detection framework in acoustically challenging environments.

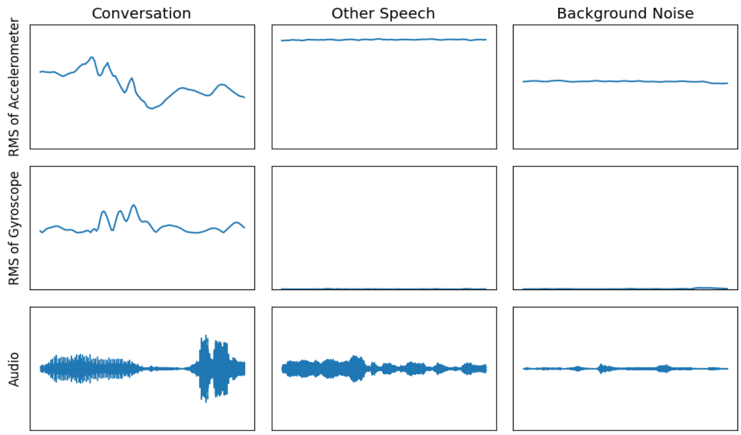

Fig. 2. Example raw acoustic and inertial data of one participant across all three classes.

5.1 Data Preprocessing

5.1.1 Audio preprocessing. As described in section 4.2, we collected the input audio at a sampling rate of 16 kHz. To obtain an audio dataset at target sampling rates of 1khz and 2khz, we downsampled the original audio to each target sampling rate. In order to maintain a consistent shape of the audio data throughout the 16kHz, 2kHz, and 1kHz datasets for model development and evaluation, we interpolated the downsampled audio through high-quality FFT-based bandlimited interpolation to achieve a data shape of the downsampled audio identical to that of the original 16kHz audio. Despite having the same data shapes across the 16kHz, 2kHz and 1kHz datasets, speech intelligibility is significantly degraded in the 2kHz and 1kHz datasets.

Within each 30-second raw audio segment, we calculated the fast Fourier Transform (FFT) using a window length of 500ms and stride of 250ms with 128 frequency bins inspired by prior works [6, 29]. This yields image-like spectrogram features in a (128x120) shape per 30-second segment, regardless of the target audio sampling rate. We normalized these FFT features before feeding them into the model as inputs.

5.1.2 IMU preprocessing. The 6-axis IMU data was collected at a sampling rate of 55Hz. We standardized the values of each IMU axis to have a mean of 0 and standard deviation of 1. Within each 30-second segment, we framed the IMU data into frames of length 2 seconds with a 1 second overlap. This corresponds to 30 IMU frames of shape (6 x 110) within a 30-second segment of data. We then extracted statistical features from the raw IMU data and converted the raw IMU data into energy per channel.

5.1.3 IMU Feature Selection. For feature selection on the IMU data, we borrow the idea of mutual information from the field of information theory. The mutual information is a measure of dependency for two discrete random variables and is defined as:

$$I(X;Y) = \Sigma_x \Sigma_y P(X,Y) log[\frac{P(X,Y)}{P(X)P(Y)}]$$

To calculate the mutual information between any given feature X and the target variable Y, where Y is the label for our three classes, we discretize the values of the feature. To achieve this, for every feature, we created a histogram with 10 bins and mapped each of the feature's values to its respective bin. Once the feature is discretized, we are able to calculate its mutual information,

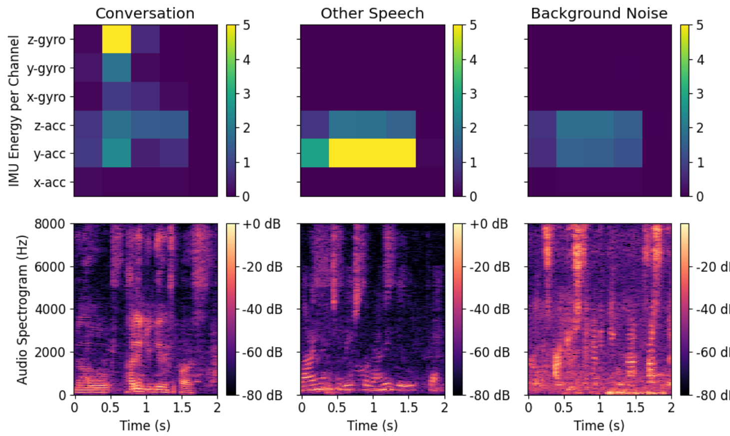

Fig. 3. Example inertial energies and acoustic spectrograms of one participant across all three classes. The audio and IMU data are synchronized.

$I(X;Y)$, with the target variable. The higher the mutual information score between a feature and the target variable, the more the feature and target variable depend on each other.

Since energy per IMU channel had a high mutual information score, we further explored using IMU energy distribution over time as a feature. We first transformed the normalized IMU signals into spectrograms by using the Short-time Fourier Transform (STFT). The STFT is a tool for transforming original time-domain signals to frequency-domain signals. The STFT of a time-series signal $x(t)$ is defined as:

$$X(\tau, f) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j2\pi ft}dt$$

where $w(t-\tau)$ is a windowing function. By taking the Fourier Transform of the original time-series signal $x(t)$ multiplied by the windowing function, we can localize in time the frequency content of the original signal. We calculated the STFT with 32 frequency bins at a time resolution of 400ms. We then calculated the energy per channel from the STFT (i.e. magnitude squared summed over all frequencies) to capture information on hand and wrist movements as they vary throughout social activities. This transformed the raw IMU signals into 30 image-like arrays of shape (6x5) per 30-second segment, as shown in Figure 3.

5.2 Audio-only Models

With data collected in section 4.2, we explored audio models to establish primary results for face-to-face conversation detection using only audio inputs. Previous works have shown that convolutional neural networks (CNNs) when applied to FFT spectrograms of acoustic data are effective at detecting the presence of foreground speech [37]. Additionally, sequence models like long short-term memory networks have demonstrated capabilities in detecting speaker turns [60].

By common consensus on the definition of face-to-face conversations, the presence of both foreground speech and turn-taking is required [14, 22, 48]. Therefore, we built upon a state-of-the-art acoustic model that incorporates both the detection of foreground speech and speaker turns into a single architecture [29]. In this acoustic model, the audio spectrograms are passed through a CNN that serves as a second feature extraction module by inferring the presence of foreground speech (Figure 4). These foreground speech embeddings, along with embeddings extracted from the original audio spectrograms, are used as input features to a LSTM network to then capture

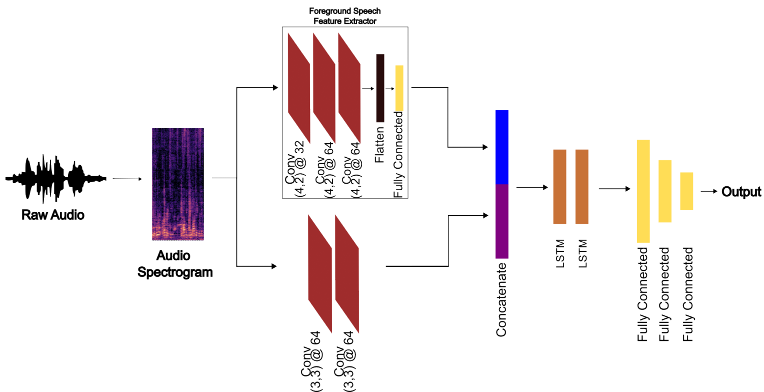

Fig. 4. Architecture of the audio classifier.

the presence of foreground speaker turn changes. Three fully connected layers follow the LSTM network to make a final prediction on the input audio.

5.3 Inertial-only Models

Using motion data, we explored two neural network frameworks to establish an initial performance on our dataset.

5.3.1 Neural Network Models. The neural networks were inspired by and built upon the following models: (1) Shallow Convolutional Neural Network with Batch Normalization (SCNNB) [26] and (2) Attend&Discriminate [1]. The architectures of both models are illustrated in Figure 5. For both models, we empirically observed better performance with IMU energy as inputs to the model compared to using raw IMU data and therefore used IMU energy frames as model inputs.

SCNNB: With the long-term vision of deploying a conversation detection model on edge devices with limited computational resources, we first experimented with utilizing a shallow, lightweight CNN. SCNNB is a network that achieves a performance on MNIST and CIFAR10 datasets comparable to deeper CNNs, such as MobileNets and VGGNet, with a shorter training time and fewer parameters. The network requires only a fraction of the time and space complexity required by larger CNNs and has motivated the use of shallow networks for HAR [52]. Therefore, this model is suited for our task and eventual deployment onto edge devices. We leveraged SCNNB containing two convolution layers to extract features within the image-like IMU energy per channel that differentiate hand, wrist, and arm movements of the three target classes.

Attend&Discriminate: The second model we explored is Attend&Discriminate, which has achieved state-of-the-art performance on public HAR datasets. The inputs are fed through a convolutional network. The extracted feature maps are then passed through a self-attention module that learns the interactions between sensor channels in the feature maps. The output feature maps are contextualized with cross-channel interactions. Then, these feature maps are passed through a recurrent neural network to capture temporal information in the sensor channels. Lastly, these sequences pass through a temporal attention module to focus on the most relevant parts of the sequence, because time-steps do not always contribute uniformly to recognizing activities.

We drew inspiration from the original Attend&Discriminate model to guide our model development. The input to the model is the energy for each channel over time, and the network learns interactions between sensor channel energies. We modified the model architecture to remove the recurrent neural network and temporal attention module, as we observed significantly lower model

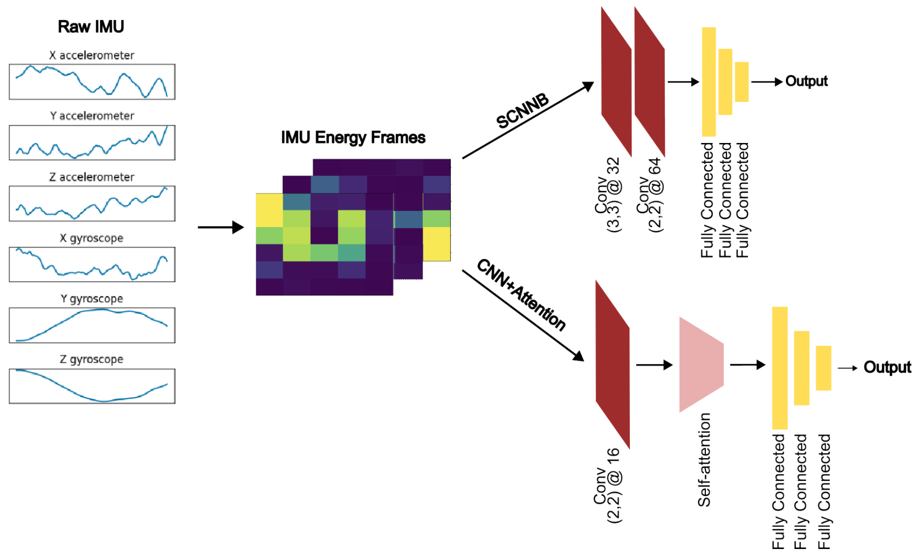

Fig. 5. Architectures of motion models explored in this work. Both models take as input IMU energy distributions over time, which has an image-like form.

accuracy with the inclusion of the sequence network and temporal attention module. We therefore call this model the *CNN+Attention* network.

5.4 Fusion Methods and Models

Standard techniques for fusing data from different sensor types include early-fusion and late-fusion [42]. In early-fusion, data from each modality is concatenated together and the concatenated data is input into the machine learning model. In late-fusion, each modality is learned independently through separate networks and the learned representations are consolidated via an aggregation operation either at: 1) representation-level or 2) score-level. Representation-level fusion can be a concatenation of each modality's embeddings or a cross-modality attention module that captures the inter-modality relationships between each sensor's representations. This fusion is followed by a single classification head for joint training of each modality's network. In score-level fusion, each network is trained separately for each modality and the predicted class probabilities per network are averaged to obtain a final class prediction. The methods are illustrated in Figure 6.

In this work, since the acoustic and inertial data have different sampling rates and are preprocessed differently, simple concatenation of the raw audio and inertial data at the input stage is not possible. Thus, we shifted our focus to late-fusion. We experimented with different methods of fusing the acoustic and inertial embeddings and predicted class probabilities. Specifically, we fused the representations of the customized acoustic model with all six inertial models at different stages in order to better understand the impacts of data fusion for conversation detection.

6 EVALUATION AND RESULTS

Our objective is to analyze and evaluate the multimodal data obtained from a smartwatch for recognizing spoken, face-to-face conversations. By leveraging the dataset collected in section 4.2, we seek to investigate the following questions:

- What degree of information does each modality provide towards conversation detection?

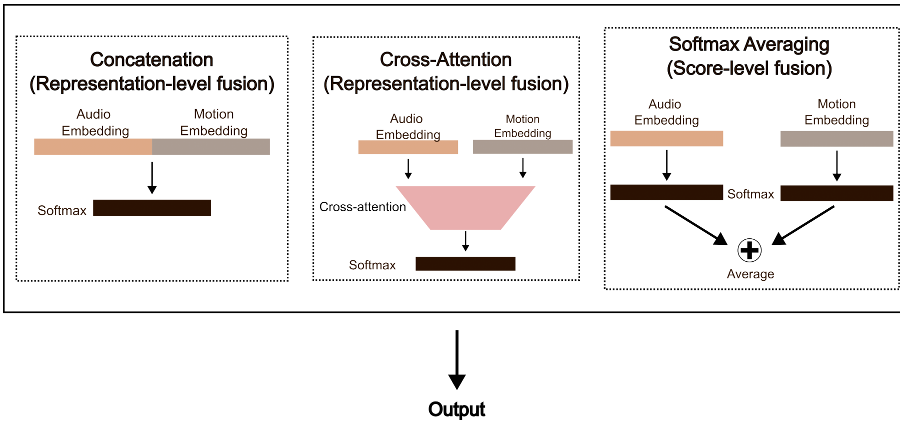

Fig. 6. An overview of fusion strategies explored in this work. Audio and inertial embeddings are first extracted from their respective networks. Representation-level fusion corresponds to combining embeddings of both modalities through concatenation or cross-attention. Score-level fusion through softmax averaging refers to training each modality's network separately and averaging the predicted probabilities.

- To what extent does the fusion of acoustic and inertial data contribute to conversation detection?
- How does a model pre-trained on data collected in a controlled *lab* setting perform on unseen data collected in *semi-naturalistic* settings?

6.1 Evaluation Setup

Conducting the study across both acoustically controlled and noisy environments and allowing participants to choose to wear the smartwatch on the wrist of their choice introduced a significant amount of variability in the acoustic environment and in data collected of participants' non-verbal communication during the activities. With all of the data collected in section 4.2, we performed three sets of evaluations to better understand the role of multimodal data for conversation detection. The three sets of evaluation are as follows: (1) evaluation in the *lab-only* setting, (2) evaluation in the *lab and semi-naturalistic* settings, and (3) evaluation in only *semi-naturalistic* settings on *lab*-trained models. Across all evaluation setups, we report the macro (unweighted) F1-score averaged across all groups used in each evaluation setup. The macro F1-score treats all target classes with equal importance, thereby removing the effects of imbalanced class distribution in the evaluation set. We obtain the 95% confidence intervals for the F1-scores by bootstrap sampling the test dataset 200 rounds.

In the *lab-only* and *lab and semi-naturalistic* evaluations, we explore how well the model generalizes across controlled and noisy settings respectively. We followed a leave-one-group-out (LOGO) cross-validation scheme in which all but one group of participants were used for training and the remaining group was used for testing. This was repeated 4 times through all combinations in the *lab-only* dataset and 11 times through all combinations in the *lab and semi-naturalistic* combined datasets due to having 4 and 7 groups in the *lab* and *semi-naturalistic* settings respectively. To evaluate how well the *lab* dataset generalizes to *semi-naturalistic* settings, in the *semi-naturalistic*-only evaluation we trained models using only *lab*-collected data and evaluated models using only data collected from the *semi-naturalistic* setting. Table 3 shows a summary of results obtained for each model across all evaluation setups.

Table 3. Average macro-F1 score for each audio-only, motion-only, and audio plus motion model with all combinations of fusion strategies evaluated across three evaluation setups on the collected dataset. L-LOGO: Training on all but one lab session and evaluating on the holdout lab session. L+SN-LOGO: training on all but one session across the lab and semi-naturalistic sessions and evaluating on the holdout session. SN: training on all lab sessions and evaluating on all semi-naturalistic sessions.

	Model	Fusion	L-LOGO	L+SN-LOGO	SN
Audio	Pure-Acoustic Model [29]	-	74.4 ± 3.3	73.0 ± 2.0	64.9 ± 2.8
Motion	SCNNB	-	51.6 ± 4.1	53.7 ± 2.4	49.0 ± 3.1
	CNN+Attention	-	49.6 ± 4.3	50.8 ± 2.5	48.7 ± 3.2
Audio+Motion	Pure-Acoustic Model + SCNNB	Softmax Averaging	67.9 ± 2.2	61.4 ± 2.3	55.2 ± 2.7
		Concatenation	77.5 ± 3.0	**78.0 ± 1.7**	67.5 ± 2.5
		Self-Attention	62.9 ± 3.8	57.9 ± 2.3	62.8 ± 2.9
	Pure-Acoustic Model CNN+Attention	Softmax Averaging	62.4 ± 3.6	65.2 ± 2.6	50.7 ± 2.6
		Concatenation	**82.0 ± 3.0**	77.2 ± 1.8	**68.1 ± 2.8**
		Self-Attention	59.6 ± 4.0	55.6 ± 2.5	55.8 ± 2.9

6.2 Evaluation Results

To assess the advantages of fusing acoustic and inertial data, we first evaluated the performance of both the acoustic and inertial models separately (Table 3 and Table 4). For acoustic-based classification, the Pure-Acoustic Model [29] achieved a macro-F1 score of 74.4±3.3% on the *lab* dataset. Among the motion models, the SCNNB model performs the best, reaching an F1 score of 53.7±2.4% on the *lab and semi-naturalistic* evaluation.

In combining acoustic and inertial data, we observe an increase in performance across all models that employ concatenation for representation-level fusion of the acoustic and inertial embeddings. However, score-level fusion and representation-level fusion through attention did not improve upon the top single-modality classifier. Fusing the embeddings extracted from the Pure-Acoustic Model for the audio data and from the CNN+Attention architecture for the inertial data achieves the best F1-score of 82.0±3.0% in evaluation in the *lab* setting, representing a 7.6%-point improvement in F1-score over the best single-modality classifier under the same evaluation setting. In *lab and semi-naturalistic* evaluation, the fused audio model and SCNNB inertial model achieves the highest F1-score of 78.0±1.7% yielding a 5.0%-point increase in F1-score compared to the audio-only classifier. For the *semi-naturalistic*-only evaluation, the fused audio and CNN+Attention inertial model again performs the best with an F1-score of 68.1±2.7%, which is 3.2%-points higher than that of the audio-only model. Furthermore, the decrease in F1-score in this evaluation setup shows the limitation of models trained entirely with data from acoustically controlled environments when evaluated on *semi-naturalistic* contexts. Since the fused audio and CNN+Attention inertial model outperforms the fused audio and SCNNB model in two of the three evaluations, we consider the Pure-Acoustic Model with CNN+Attention through Concatenation to be the best performing multimodal classifier.

This improvement in multimodal model performance comes with only a 0.4% increase in number of model parameters compared to the best audio-only classifier. The classifiers have 2.8K, 763.2K, and 766.5K parameters for the inertial-only (CNN+Attention), audio-only, and multimodal classifiers respectively. This highlights the lightweight manner in which gestures and body movements captured by inertial data can be effectively incorporated for conversation detection.

Confusion matrices comparing the performance of the single modality classifiers composing the top multimodal classifier and the multimodal classifier itself are shown in Figure 7. Per-group performance for the audio (Pure-Acoustic Model), inertial (CNN+Attention), and multimodal (Pure-Acoustic Model with CNN+Attention through Concatenation) models are shown in Figure 8. The

Table 4. Macro precision and recall. P: precision. R: recall.

Modality	Model	L-LOGO		L+SN-LOGO		SN	
		P	R	P	R	P	R
Audio	Pure-Acoustic Model	77.6 ± 3.4	74.2 ± 3.0	73.5 ± 2.1	73.4 ± 1.9	65.3 ± 2.4	64.8 ± 2.8
IMU	SCNNB	55.8 ± 5.0	51.8 ± 3.5	58.9 ± 2.7	53.5 ± 2.2	50.2 ± 3.4	48.8 ± 3.0
	CNN+Attention	56.0 ± 5.4	49.8 ± 3.5	59.7 ± 3.0	51.3± 2.1	48.9 ± 3.0	48.9 ± 3.1
Audio+IMU	Pure-Acoustic Model + SCNNB	79.4 ± 3.6	77.3 ± 3.0	**78.2 ± 1.8**	**78.3 ± 1.7**	69.9 ± 2.6	67.6 ± 2.6
	Pure-Acoustic Model + CNN+Attention	**82.4 ± 3.1**	**82.7 ± 2.7**	77.7 ± 2.0	77.0 ± 1.8	**71.1 ± 2.6**	**68.2 ± 2.5**

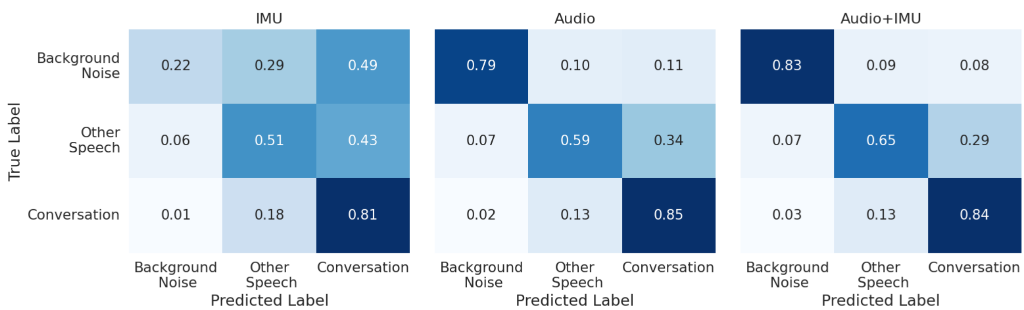

Fig. 7. Confusion matrices for inertial-only (left), audio-only (center), and audio-inertial frameworks (right).

lab groups have an average F1-score of 80.0±3.7% while the *semi-naturalistic* groups have an average F1-score of 75.4±3.4% , though group 7 with the highest F1-score of 82.0±3.2% comes from the *semi-naturalistic* setting. On the other hand, groups 8 and 11 have the worst F1-scores of 62.0±3.8% and 65.6±3.8% respectively. Upon examination of these two groups, we found that the declined performance could be due to two factors. First, both smartwatch users in group 8 and one smartwatch user in group 11 wore the watch on their dominant hand. Since a significant majority of inertial data (17 out of 22 participants) came from individuals' non-dominant hand, the model may have learned to better leverage inertial data generated by participants' non-dominant hand. In watching the recorded videos of each group, we observed that participants gestured less frequently and dramatically with their non-dominant hand compared to their dominant hand. Second, groups 8 and 11 were the only groups whose data collection was outdoors. Acoustic and motion artifacts unique to the outdoors setting in their small sample size could have degraded model performance. Additional data collection outdoors and of participants who choose to wear smartwatches on their dominant hands could help mitigate these issues in the future.

7 DISCUSSION

In this section, we discuss additional evaluations performed across window lengths, activity contexts, audio sampling rates, and dataset environments and the tradeoff between ecological validity and participant handedness during the data collection study.

7.1 Frame Sensitivity Analysis

Frame sizes in HAR impact classification granularity, feature extraction, and model performance. Therefore, we gauged the impact of overall window length and IMU frame size for our novel approach on social interaction analysis in busy environments.

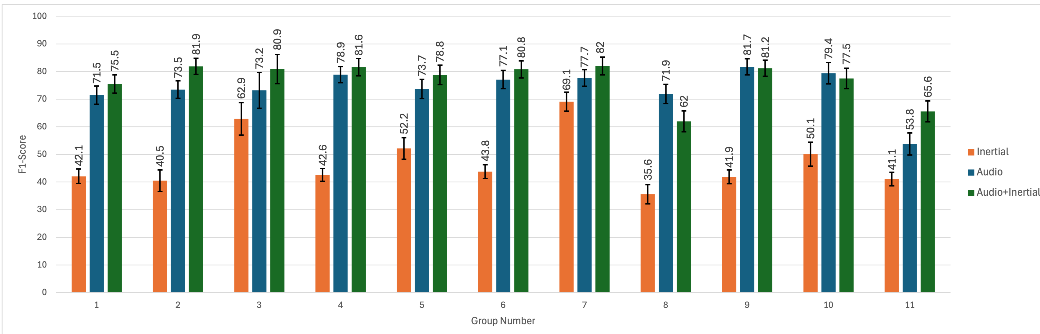

Fig. 8. A comparison of LOGO evaluation results (macro F1-score) across audio, inertial, and multimodal classifiers. The multimodal framework improves upon any single-modality classifier in all but three groups.

7.1.1 Overall Frame Sensitivity. Given the dynamic fluctuations characteristic of busy settings, we investigated the model's performance across window lengths in 10-second intervals spanning from 10 to 30 seconds. Again, we evaluate the model with a LOGO cross-validation on *lab-only* and *lab and semi-naturalistic* data and with an evaluation on *semi-naturalistic* data upon *lab-only* training.

Figure 10 shows that multimodal performance improves as window length increases, with peak performance at a window length of 30 seconds. This is in line with previous works that have found a 30 second window to provide the best tradeoff between classification accuracy and robustness [4, 12, 29, 44]. Although longer windows give better model performance, it reduces the granularity of conversation detection as each window can only be assigned to a single event class.

7.1.2 IMU Frame Size Sensitivity. We further assess the impact of IMU frame size in conversation detection with the overall prediction window length fixed at 30 seconds. Towards this goal, we evaluate the inertial-only (CNN+Attention) and multimodal (Pure-Acoustic Model with CNN+Attention through Concatenation) frameworks on IMU frame sizes varying from 1 to 10 seconds in 1 second increments. We use a LOGO evaluation on the *lab and semi-naturalistic* data and report the macro F1-score.

Both inertial-only and acoustic-inertial models perform relatively consistently through the various IMU frame sizes (Figure 10). In the inertial-only model, model performance trends upward with larger frame lengths, achieving a maximum macro F1-score of 51.9±2.6% at a frame size of 9s. In the acoustic-inertial model, model performance oscillates across IMU frame lengths and reaches a maximum macro F1-score of 77.2±1.8% at a frame length of 2s. As we are primarily focused on the multimodal model, we proceeded with a 2-second IMU frame length for this classification task.

7.2 Multimodality Benefits by Context

To further understand specific contexts that benefit most from additional non-verbal communication in conversation sensing, we evaluate the single-modality and multimodal classifiers by activity type on the combined *lab and semi-naturalistic* dataset. The specific activity types we consider are: 1) regular conversation, 2) conversation while eating, 3) reading out loud, 4) watching videos, and 5) music in background. For each activity type, we evaluate using LOGO cross validation the top single-modality and multimodal classifiers on all data segments that contain the target activity context and repeat the process for each activity type. For instance, for *music in background*, the evaluation dataset is all data segments across all groups of the study that contained background music. These activity types are not mutually exclusive, except between *regular conversation* and *conversation while eating.*

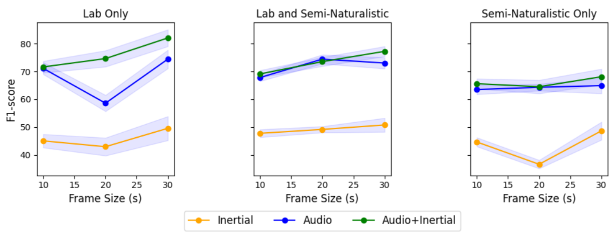

Fig. 9. A comparison of macro F1-scores across all three evaluation setups for acoustic, inertial and multimodal classifiers with window lengths varying from 10 to 30 seconds.

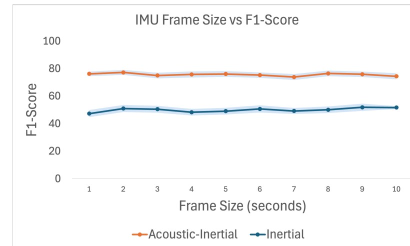

Fig. 10. A comparison of *lab+semi-naturalistic* LOGO evaluation results across inertial and multimodal classifiers with varying IMU frame sizes.

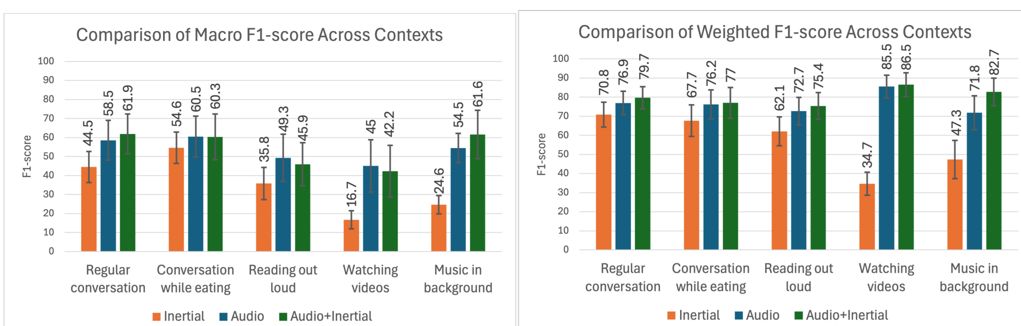

Fig. 11. Macro (left) and weighted (right) F1-scores across conversation activity contexts.

In this context-based analysis, there are significantly more class imbalances in the evaluation datasets. For instance, there are likely to be fewer instances of the *conversation* class while watching videos. Therefore, we report both macro and weighted F1 scores. Macro F1-score evaluates model performance independently of the class distributions in the evaluation set. However, this makes the metric sensitive to rare classes and can be significantly influenced by rare labels. The weighted F1-score addresses this by taking into consideration class distributions.

Across all activity contexts, the joint acoustic-inertial modality improves upon performance (weighted F1-score) of any single modality classifier. The addition of the inertial modality is most beneficial to the *music in background* context, increasing the absolute mean value of the weighted F1-score of the audio-only model by 10.9%. The benefit of the inertial modality makes sense, as people often tap to the beat of the music or move in other ways that are distinct from other social activities. This finding especially can be leveraged for analyzing conversations in settings that often contain background music, such as dining at a restaurant or gatherings at a bar. Overall, this evaluation highlights the effectiveness of joint acoustic-inertial sensing compared to single-modality sensing across a variety of contexts around which conversations are commonly centered.

7.3 Audio Privacy Benefits of Multimodality

With the addition of inertial data to a traditionally audio-only classification problem, we explored whether the information contained in inertial data can be leveraged to reduce the amount of information required in the audio data without sacrificing task performance. The minimum sampling rate for intelligible speech is commonly 16kHz, since most speech occurs below 8kHz [49]. Speech

Table 5. Average macro-F1 score for audio and audio-inertial models with the audio sampled at three different sampling rates (16kHz, 2kHz, 1kHz). L-LOGO: Training on all but one *lab* sessions and evaluating on the holdout lab session. SN-LOGO: training on all but one session across the *lab and semi-naturalistic* sessions and evaluating on the holdout session. SN: training on all *lab* sessions and evaluating on all *semi-naturalistic* sessions.

Modality	Model	Frequency (kHz)	L-LOGO	L+SN-LOGO	SN
Audio	Pure-Acoustic Model	16	74.4 ± 3.3	73.0 ± 2.0	64.9 ± 2.8
		2	73.2 ± 3.4	74.1 ± 2.0	60.5 ± 1.6
		1	69.0 ± 2.3	72.3 ± 3.7	54.3 ± 2.5
Audio+Motion	Pure-Acoustic Model + SCNNB	16	77.5 ± 3.0	78.0 ± 1.7	67.5 ± 2.5
		2	69.7 ± 4.0	75.5 ± 2.0	64.5 ± 2.3
		1	71.9 ± 3.8	73.8 ± 1.8	61.7 ± 2.6
	Pure-Acoustic Model + CNN+Attention	16	82.0 ± 3.0	77.2 ± 1.8	68.1 ± 2.8
		2	73.6 ± 3.2	75.7 ± 2.0	64.8 ± 2.7
		1	73.2 ± 3.7	74.6 ± 1.8	62.2 ± 2.5

intelligibility decreases as audio sampling rates decrease below 16kHz, with one study showing a significant drop in intelligibility at a sampling rate of 1kHz [35]. Though more intelligible, audio sampled at 16kHz comes with privacy concerns about recording the content of a user's speech. In contrast, inertial data sampled at 50Hz is less privacy sensitive than audio data [31, 56]. Therefore, we investigated whether non-verbal communication captured through inertial data can supplement sub-sampled audio for conversation detection to increase privacy-preservation for the user.

For this exploration, we created a sub-sampled audio dataset at 2kHz and 1kHz from the original audio dataset collected at 16kHz following the process described in section 5.1. We trained and evaluated the customized audio-only model and the top two combined audio-inertial models on audio sampled at 16kHz, 2kHz, and 1kHz. We report the macro F1-score for the same three evaluation setups described in section 6.1, which are: (1) evaluation in the *lab*-only setting, (2) evaluation in the *lab and semi-naturalistic* settings, and (3) evaluation in only the *semi-naturalistic* settings using *lab* trained models.

As expected, the audio-only model's performance decreases across all three evaluation setups as the sampling rate decreases from 16kHz to 1kHz. In contrast, the top performing multimodal model (Pure-Acoustic Model and CNN+Attention) outperforms the audio-only model across all nine combinations of frequency and evaluation scenarios. While the audio-inertial model performance still decreases as audio quality decreases, the drop in model performance is not as significant as the performance drop in the audio-only model. Therefore, the addition of the inertial modality shows that IMU data can effectively supplement information lost in downgraded audio for detecting conversations. The combined audio and inertial framework is more robust to low-quality audio, and this finding can be leveraged to perform audio sensing at a lower sampling rate to maintain user privacy.

7.4 Validating the Challenge of Dynamic Environments for Conversation Detection

To better understand the model developed using our dataset in context with models developed on previous acoustic smartwatch datasets for conversations, we evaluated the trained model presented by Liang *et al.* [29] on the audio in our collected dataset. Their study collected a *semi-naturalistic* dataset with 32 hours of audio recorded in 18 homes while all household members engaged in a set of scripted activities and a *free-living* dataset with 45 hours of audio recorded by 4 individuals in real-world settings without any activity constraints. Both datasets were recorded using smartwatch microphones as well. Notably, only their *semi-naturalistic* dataset from home environments was

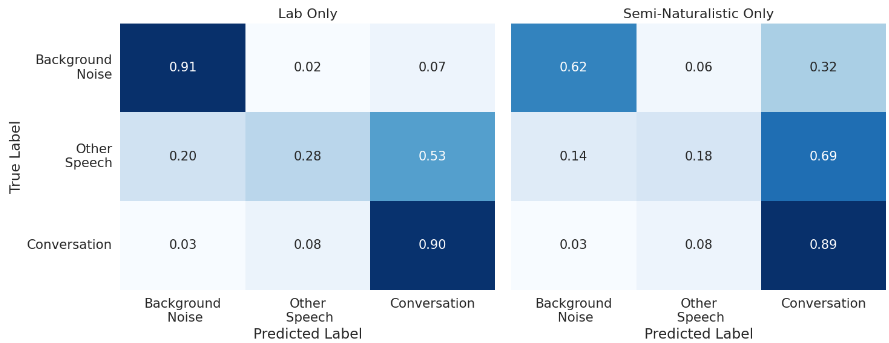

Fig. 12. Confusion matrices from inference of the pre-trained audio model on our collected dataset. Left: Inference on data collected in *lab* settings (groups 1-4). Right: Inference on data collected in *semi-naturalistic* settings (groups 5-11).

used to train their acoustic model while both their *semi-naturalistic* and *free-living* datasets were used for model evaluation. On their *semi-naturalistic* dataset, their model achieved a macro-F1 score of 76.2% and on their *free-living* dataset, they achieved a macro-F1 score of 89.2%.

We leverage their model for inference only on both our *lab*-only and *semi-naturalistic*-only datasets. As seen in Figure 12, their pre-trained model performs better on our *lab* dataset (macro F1-score 68.7%) than our *semi-naturalistic* dataset (macro F1-score 52.3%). This gap in performance between their datasets and our datasets emphasizes the acoustic difficulty of the environments in which we collected our datasets.

Using their pre-trained model, the *other speech* class is significantly confused with the *conversation* class across both our *lab* and *semi-naturalistic* datasets. In our *semi-naturalistic* dataset, conversations from passersby in the environment also resulted in *background noise* confused with *conversation*. Overall, this dataset comparison demonstrates a domain shift in the data where the training distribution (i.e., their *semi-naturalistic* dataset from quieter home environments) differs from the test distribution (i.e., our *semi-naturalistic* dataset from public, noisy environments). This again highlights the uniqueness of our dataset and investigation of conversation detection in dynamic environments, which has been unexplored by previous datasets and studies.

7.5 Participant Handedness and Data Collection

As previously discussed, to increase the ecological validity of the data collection study, participants were free to wear the smartwatch on either wrist. As seen in Table 2, an overwhelming majority of participants were right-hand dominant and chose to wear the watch on their non-dominant left hand. However, during the study, it was observed that participants primarily gestured with their dominant hand, which aligns with previous studies on the relationship between gesturing and handedness [9]. Therefore, many participant gestures are not captured in the recorded inertial data. While the multimodal architecture is already an improvement from the baseline single-modality classifiers, inertial data from participants' dominant hand could help further clarify social activities, especially in noisy settings, with greater accuracy.

8 MODEL DEPLOYMENT

As previously mentioned, we used one Fossil Gen 4 smartwatch and one Fossil Gen 5 smartwatch as data collection devices. The Gen 4 smartwatch has 512MB RAM and 4GB storage while the Gen 5 smartwatch has 1GB RAM and 8GB storage. Compared to smartphones or other edge devices like Raspberry Pis, smartwatches have significantly fewer computational resources. Despite these computational limitations of smartwatches, we demonstrate that our conversation detection model can deploy to smartwatches in this section. As will be discussed in section 8.2, we focus our model deployment discussion on the Pure-Acoustic Model with SCNNB architecture after discovering hardware limitations in the smartwatches to support the Pure-Acoustic Model with CNN+Attention model architecture.

8.1 Model Optimization

To facilitate model deployment, we optimized the model through quantization-aware training and weight pruning. With quantization-aware training, we lower the precision of model parameters from 32-bit float representations to 8-bit integer representations by introducing quantization effects during model training such that the trained model is more robust to the loss in weight precision. We additionally prune 50% of the parameters per layer to remove insignificant parameters and obtain a sparser model. Rather than training from scratch, we fine-tuned the pre-trained weights of the Pure-Acoustic Model with SCNNB model. For comparison to the Pure-Acoustic Model, we also optimize the audio-only model in the same manner. We evaluate the optimized model performance with the *lab and semi-naturalistic* LOGO evaluation setup and show the results in Table 6. We observe that the joint audio-inertial optimized model achieves a performance similar to its pre-optimization performance and still outperforms the optimized, audio-only model.

8.2 Deployment to Smartwatches

We converted both optimized models to TensorFlow Lite (TFLite), a data format that allows models to run on edge devices. We then developed an Android application using Java to load and invoke the model on the smartwatches. During this process, we discovered that some operations required for the attention mechanism are not supported by the hardware in the Fossil Gen 4 and 5 smartwatches. Therefore, we focused on deployment of the optimized Pure-Acoustic Model with SCNNB model as the joint audio-inertial model. Additionally, while the audio-inertial model can run on both the Fossil Gen 4 and 5 smartwatches, we focus our deployment analysis on the Gen 5 smartwatch as it is newer than the Gen 4 smartwatch (2019 vs 2018) and has double the RAM.

We profile the inference time of both the audio-only and joint audio-inertial models to understand the potential real-time applications our system. We summarize the TFLite size and average inference times of both models while running on the Fossil Gen 5 smartwatch in Table 6. The average inference time is measured across 10 successive invocations of the model on the smartwatch.

As many smartwatches have been released since the Fossil Gen 5 smartwatch in 2019, we also profiled the model's inference time on a newer smartwatch, the Google Pixel Watch 2 released in 2023. With 2GB RAM, the average inference time of the joint audio-inertial model is reduced by over a factor of two down to 400ms. Furthermore, this smartwatch hardware supports the Pure-Acoustic Model with CNN+Attention model architecture with a similar runtime to the Pure-Acoustic Model with SCNNB model.

8.3 Cost-Benefit Analysis

We conducted a cost-benefit analysis on model performance and smartwatch battery life of the Fossil Gen 5 as a function of audio and IMU sampling rates. Although for audio privacy measures we downsampled the audio to 1kHz, the Fossil smartwatches used in data collection are limited to 4kHz as the lowest microphone sampling rate. We compare the results of audio and IMU sampling rate combinations in Table 7. We report the F1-score of the Pure-Acoustic+SCNNB model. As observed in the table, the additional IMU modality in the joint audio-inertial model provides a statistically significant improvement to the detection of conversations compared to the audio-only model.

Table 6. Comparison of the performance, size, and inference time on a Fossil Gen 5 smartwatch of the optimized audio-only and audio-inertial models.

Model	F1-score	TFLite Size (kB)	Inference Time on Smartwatch (ms)
Optimized Pure-Acoustic Model	74.6 ± 2.0	831	953.6
Optimized Pure-Acoustic Model + SCNNB	77.3 ± 1.8	857	972.5

Table 7. We profile the battery life of a Fossil Gen 5 smartwatch and model performance with varying audio and IMU sampling rates. Battery life is measured by the duration of data collection on the smartwatch using our data collection app until the battery is fully exhausted from one single, full charge. The IMU-only model is the SCNNB network and the joint audio-IMU model is the Pure-Acoustic Model with SCNNB model. We report the F1-score of the models in the L+SN-LOGO evaluation scheme.

Modality	Audio Sampling Rate (kHz)	IMU Sampling Rate (Hz)	Battery Life	Model Performance (F1 score)
Audio-only	16	-	5hrs 29min	73.0 ± 2.0
	4	-	6hrs 54min	74.6 ± 2.0
IMU-only	-	50	4hrs 26min	53.6 ± 2.3
	-	25	5hrs 28min	52.9 ± 2.5
	-	10	7hrs 6min	49.6 ± 4.3
Audio+IMU	16	50	3hrs 44min	78.0 ± 1.7
	16	25	4hrs 18min	75.9 ± 2.0
	16	10	5hrs 49min	72.6 ± 2.2
	4	50	4hrs 55min	78.0 ± 2.0
	4	25	5hrs 45min	77.2 ± 1.8
	4	10	6hrs 40min	76.5 ± 2.0

However, we note that this benefit comes at a cost of 1.75 hours of reduced smartwatch battery life. Interestingly, we note that the model performs better with 4kHz audio compared to 16kHz audio. In listening to the audio downsampled to 4kHz, we hypothesize this is due to foreground voices still being audible and discernible but background speech becoming more distorted at 4kHz.

9 APPLICATIONS

Practical recognition of social interactions will enable a wide range of new applications including, but not limited to, organizational behavior, health and wellness, and augmented reality domains. In this section, we further expand on applications of automatic conversation detection and discuss the extent to which our framework is suitable for these applications in light of our results.

- **Team Dynamics** The proposed system offers insights into team dynamics. Inclusive team dialogues, characterized by different team members contributing in succession, correlate positively to better team skill use and task strategy and lead to improved overall performance [18]. Therefore, our proposed system can identify the extremes of team discussions, whether someone is dominating the conversation or not speaking at all, and can allow teams to review their communication, coordination, and cohesion.

- **Loneliness and Social Isolation** Social isolation is as significant of a risk factor for health outcomes as traditional risk factors such as obesity [41]. After medical events such as experiencing a stroke, individuals' social networks decline and become less diverse [21]. Therefore, the proposed social interaction sensor can allow physicians and care providers to better support and understand the relationship between patients who have experienced such medical events and patient outcomes.
- **Social Diary** Detecting social interactions allows individuals to maintain logs of their daily social interactions. The user can capture information, such as time and duration of their daily interactions, giving users a comprehensive view of their social activities. Longitudinally, these social diaries can increase speakers' self-awareness of their social interactions and identify potential patterns of isolation.

At its current performance, our approach to sensing social interactions in noisy environments is suitable for these applications to a certain extent. To understand broad trends of team dynamics or loneliness and social isolation, the current framework could be acceptable. For other applications that require specific precision and recall to detect subtle nuances in conversation dynamics for enabling high-fidelity health analyses for instance, our system may need to be fine-tuned to the specific application context or improved with additional features discussed in the following section. Overall, though, our system represents a significant first step towards realizing these applications.

10 LIMITATIONS AND FUTURE WORK

While our work demonstrates the capabilities of smartwatch social sensing in naturalistic noisy settings, it is important to highlight its limitations and discuss future opportunities. First, though we evaluated our framework on a *semi-naturalistic* dataset collected in real-world acoustic settings, we did not evaluate the framework on an *in-the-wild* dataset where participants were unsupervised in their activities. In real-world settings, people can multitask, such as walking or driving, while engaged in a conversation. These activities that overlap with conversations, especially these activities that also have accompanying hand, arm, or wrist gestures, may create confusion with conversation-related gestures. Secondly, while we collected data from 35 participants with diverse gender and cultural representation, all participants were between the ages of 20-30 years old. Additional participants with more diversity in age would enhance the external validity of our results, since non-verbal and verbal communication vary across ages [15]. We intend to address these limitations in future work to improve upon the current system.

Additionally, while speech processing tools alone are not currently suitable for social interaction analysis, they can extract information that can assist in further characterizing conversations. For instance, *pyannote.audio* can perform speaker diarization and detect overlapping speech. By segmenting audio according to who spoke when, the information gained through speaker diarization could improve classification of the *conversation* and *other speech* classes. However, it is important to note that many speech processing tools have primarily been developed on datasets from controlled environments such as LibriSpeech [40] and Switchboard [17], which come from audiobooks and telephone calls respectively. Therefore, these tools alone have limited applicability to detecting in-person conversations in acoustically challenging environments. However, coupled with our framework, these tools can contribute additional analyses such as characterization of speech overlaps that are of interest to the field of conversation analysis [20].

11 CONCLUSION

We present the first joint acoustic-inertial sensing framework using off-the-shelf smartwatches for recognizing conversations. We demonstrate the benefits of inertial data in capturing non-verbal behaviors during in-person communication to aid acoustic sensing. To validate this framework, we collected two datasets: (1) a *lab* dataset with 11 participants split into 4 groups performing 5 supervised group activities and (2) a *semi-naturalistic* dataset with 24 participants split into 7 groups performing the same group activities in acoustically challenging environments. Through a broad set of evaluations, we show the advantages of multimodal sensing for conversation detection in acoustically-challenging environments, which has been previously unexplored. Furthermore, we demonstrate the advantages of inertial data in aiding model performance across activity contexts and low-quality audio. This work advances the development of high-performing conversation detection systems by utilizing everyday wrist-worn devices to analyze both acoustic and inertial data. Our framework opens the door for building future systems for more fine-grained analyses of social dynamics for applications towards individual well-being, organizational behavior and more.

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] Apple. 2023. AirPods redefine the personal audio experience. https://www.apple.com/newsroom/2023/06/airpods-redefine-the-personal-audio-experience/ Accessed: 2023-12-19.

[3] Rummana Bari, Roy J. Adams, Md. Mahbubur Rahman, Megan Battles Parsons, Eugene H. Buder, and Santosh Kumar. 2018. rConverse: Moment by Moment Conversation Detection Using a Mobile Respiration Sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 2 (mar 2018), 27 pages. https://doi.org/10.1145/3191734

[4] Rummana Bari, Md. Mahbubur Rahman, Nazir Saleheen, Megan Battles Parsons, Eugene H. Buder, and Santosh Kumar. 2020. Automated Detection of Stressful Conversations Using Wearable Physiological and Inertial Sensors. 4, 4, Article 117 (dec 2020), 23 pages. https://doi.org/10.1145/3432210

[5] Janet Bavelas and Jennifer Gerwing. 2007. *Conversational hand gestures and facial displays in face-to-face dialogue.* 283–308.

[6] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: combining audio and motion sensing for gesture recognition on smartwatches. In *2019 ISWC* (London, United Kingdom) *(ISWC '19)*. ACM, New York, NY, USA, 10–19. https://doi.org/10.1145/3341163.3347735

[7] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 42 (jul 2022), 28 pages. https://doi.org/10.1145/3534582

[8] Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

[9] Esra Nur Çatak, Alper Açık, and Tilbe Göksun. 2018. The relationship between handedness and valence: A gesture study. *Quarterly Journal of Experimental Psychology* 71, 12 (2018), 2615–2626. https://doi.org/10.1177/1747021817750110 arXiv:https://doi.org/10.1177/1747021817750110 PMID: 29355469.

[10] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. 2010. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLOS ONE* 5, 7 (07 2010), 1–9. https://doi.org/10.1371/journal.pone.0011596

[11] Metod Celestina, Jan Hrovat, and Chucri A. Kardous. 2018. Smartphone-based sound level measurement apps: Evaluation of compliance with international sound level meter standards. *Applied Acoustics* 139 (2018), 119–128. https://doi.org/10.1016/j.apacoust.2018.04.011

[12] Tanzeem Choudhury and Alex Pentland. 2003. Sensing and Modeling Human Networks. (01 2003).

[13] International Noise Awareness Day. 2024. https://noiseawareness.org/info-center/common-noise-levels/ Accessed: 2024-09-05.

[14] Susan Kay Donaldson. 1979. One Kind of Speech Act: How Do We Know When We ?Re Conversing?? *Semiotica* 28, 3-4 (1979). https://doi.org/10.1515/semi.1979.28.3-4.259

[15] R.S. Feldman and J.M. Tyler. 2006. *Factoring in age: Nonverbal communication across the life span.* 181–200. https://doi.org/10.4135/9781412976152.n10

[16] Georgi Gerganov. 2023. whisper.cpp. https://github.com/ggerganov/whisper.cpp.

[17] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *1992 ICASSP* (San Francisco, California) *(ICASSP'92)*. IEEE Computer Society, USA, 517–520.

[18] Ki-Won Haan, Christoph Riedl, and Anita Woolley. 2021. Discovering Where We Excel: How Inclusive Turn-Taking in Conversation Improves Team Performance. In *2021 ICMI* (Montreal, QC, Canada) *(ICMI '21 Companion)*. ACM, New

York, NY, USA, 278–283. https://doi.org/10.1145/3461615.3485417

[19] Jay Hall and W. H. Watson. 1970. The Effects of a Normative Intervention on Group Decision-Making Performance. *Human Relations* 23, 4 (1970), 299–317. https://doi.org/10.1177/001872677002300404

[20] Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38 (10 2010), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

[21] Katerina Hilari and Sarah Northcott. 2016. "Struggling to stay connected": comparing the social relationships of healthy older people and people with stroke and aphasia. *Aphasiology* (08 2016), 1–14. https://doi.org/10.1080/02687038.2016.1218436

[22] Judith Holler, Kobin H. Kendrick, Marisa Casillas, and Stephen C. Levinson (Eds.). 2016. *Turn-Taking in Human Communicative Interaction*. Frontiers Media S.A., Lausanne. https://eprints.whiterose.ac.uk/116190/

[23] Chiao-Yin Hsiao, Wan-Rong Jih, and Jane Hsu. 2012. Recognizing Continuous Social Engagement Level in Dyadic Conversation by Using Turn-taking and Speech Emotion Patterns.

[24] Hyunchoong Kim, Jonghoon Shin, Soohwan Kim, Yohan Ko, Kyoungwoo Lee, Hojung Cha, Seong Il Hahm, and Taejun Kwon. 2017. Collaborative classification for daily activity recognition with a smartwatch *(2016 IEEE International Conference on Systems, Man, and Cybernetics)*. Institute of Electrical and Electronics Engineers Inc., 3707–3712. https://doi.org/10.1109/SMC.2016.7844810

[25] Youngki Lee, Chulhong Min, Chanyou Hwang, Jaeung Lee, Inseok Hwang, Younghyun Ju, Chungkuk Yoo, Miri Moon, Uichin Lee, and Junehwa Song. 2013. SocioPhone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. *MobiSys 2013*, 375–388. https://doi.org/10.1145/2462456.2465426

[26] Fangyuan Lei, Xun Liu, Qingyun Dai, and Bingo Ling. 2020. Shallow convolutional neural network for image classification. *SN Applied Sciences* 2 (01 2020). https://doi.org/10.1007/s42452-019-1903-4

[27] Qiang Li, Shanshan Chen, and John Stankovic. 2013. Multi-modal in-person interaction monitoring using smartphone and on-body sensors. *2013 IEEE International Conference on Body Sensor Networks, BSN 2013*, 1–6. https://doi.org/10.1109/BSN.2013.6575509

[28] Dawei Liang, Guihong Li, Rebecca Adaimi, Radu Marculescu, and Edison Thomaz. 2022. AudioIMU: Enhancing Inertial Sensing-Based Activity Recognition with Acoustic Models. In *2022 ISWC* (Cambridge, United Kingdom) *(ISWC '22)*. ACM, New York, NY, USA, 44–48. https://doi.org/10.1145/3544794.3558471

[29] Dawei Liang, Alice Zhang, and Edison Thomaz. 2023. Automated Face-To-Face Conversation Detection on a Commodity Smartwatch with Acoustic Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 109 (Sept 2023), 29 pages. https://doi.org/10.1145/3610882

[30] Chengwen Luo and Mun Chan. 2013. SocialWeaver: Collaborative inference of human conversation networks using smartphones. (11 2013), 14. https://doi.org/10.1145/2517351.2517353

[31] Aleksandar Matic, Venet Osmani, and Oscar Mayora. 2013. *Automatic Sensing of Speech Activity and Correlation with Mood Changes*. 195–205. https://doi.org/10.1007/978-3-642-32538-0_9

[32] David Matsumoto and Hyisung Hwang. 2016. *The cultural bases of nonverbal communication*. 77–101. https://doi.org/10.1037/14669-004

[33] Albert Mehrabian. 1972. *Nonverbal Communication*. Routledge.

[34] Emiliano Miluzzo, Cory T. Cornelius, Ashwin Ramaswamy, Tanzeem Choudhury, Zhigang Liu, and Andrew T. Campbell. 2010. Darwin Phones: The Evolution of Sensing and Inference on Mobile Phones. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services* (San Francisco, California, USA) *(MobiSys '10)*. ACM, New York, NY, USA, 5–20. https://doi.org/10.1145/1814433.1814437

[35] 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 (sep 2022), 19 pages. https://doi.org/10.1145/3550284

[36] Debbie S Moskowitz and Simon N Young. 2006. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *J. Psychiatry Neurosci.* 31, 1 (Jan. 2006), 13–20.

[37] Amrutha Nadarajan, Krishna Somandepalli, and Shrikanth S. Narayanan. 2019. Speaker Agnostic Foreground Speech Detection from Audio Recordings in Workplace Settings from Wearable Recorders. *2019 ICASSP* (2019), 6765–6769. https://api.semanticscholar.org/CorpusID:145905315

[38] Daniel Olguin Olguin, Benjamin N. Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. 2009. Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2009), 43–55. https://doi.org/10.1109/TSMCB.2008.2006638

[39] World Health Organization. 2021. Social isolation and loneliness among older people: advocacy brief. https://www.decadeofhealthyageing.org/find-knowledge/resources/publications/un-decade-of-healthy-ageing-advocacy-brief-social-isolation-and-loneliness-among-older-people Accessed: 2023-12-19.

[40] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 ICASSP*. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[41] Matthew Pantell, David Rehkopf, Douglas Jutte, S Leonard Syme, John Balmes, and Nancy Adler. 2013. Social isolation: a predictor of mortality comparable to traditional clinical risk factors. *Am. J. Public Health* 103, 11 (Nov. 2013), 2056–2062.

[42] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. 2023. Effective Techniques for Multimodal Data Fusion: A Comparative Analysis. *Sensors* 23, 5 (2023). https://doi.org/10.3390/s23052381

[43] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/ARXIV.2212.04356

[44] Md Rahman, Amin Ali, Kurt Plarre, Mustafa al'Absi, Emre Ertin, and Santosh Kumar. 2011. MConverse: Inferring conversation episodes from respiratory measurements collected in the field. *Proceedings - Wireless Health 2011, WH'11*, 10. https://doi.org/10.1145/2077546.2077557

[45] Mary Redmayne. 2017. Where's Your Phone? A Survey of Where Women Aged 15-40 Carry Their Smartphone and Related Risk Perception: A Survey and Pilot Study. *PLOS ONE* 12 (01 2017), e0167996. https://doi.org/10.1371/journal.pone.0167996

[46] Harry T. Reis and Ladd Wheeler. 1991. *Studying Social Interaction with the Rochester Interaction Record.* Elsevier, 269–318. https://doi.org/10.1016/s0065-2601(08)60332-9

[47] Sujiya S and Dr.Chandra E. 2017. A Review on Speaker Recognition. *International Journal of Engineering and Technology* 9 (06 2017), 1592–1598. https://doi.org/10.21817/ijet/2017/v9i3/170903513

[48] Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A Simple Systematic for the Organisation of Turn Taking in Conversation. *Language* 50 (12 1974), 696–735. https://doi.org/10.2307/412243

[49] Florian Schiel, Christoph Draxler, Angela Baumann, Tania Ellbogen, and Alex T. Steffen. 2012. The Production of Speech Corpora. https://api.semanticscholar.org/CorpusID:61111457

[50] Nabeel Siddiqui and Rosa Chan. 2020. Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist. *PLOS ONE* 15 (01 2020), e0227039. https://doi.org/10.1371/journal.pone.0227039

[51] NIOSH Hearing Loss Prevention Team. 2019. *NIOSH Sound Level Meter Application (app) for iOS devices.*

[52] Dipanwita Thakur and Suparna Biswas. 2021. Feature fusion using deep learning for smartphone based human activity recognition. *International Journal of Information Technology* 13 (06 2021). https://doi.org/10.1007/s41870-021-00719-6

[53] Yonatan Vaizman, Katherine Ellis, Gert Lanckriet, and Nadir Weibel. 2018. ExtraSensory App: Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior. In *2018 CHI (CHI '18)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174128

[54] 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 (01 2018), 1–22. https://doi.org/10.1145/3161192

[55] Martin Warren. 2006. *Features of Naturalness in Conversation.* https://doi.org/10.1075/pbns.152

[56] Gary Weiss, Jeffrey Lockhart, Tony Pulickal, Paul McHugh, Isaac Ronan, and Jessica Timko. 2016. Actitracker: A Smartphone-Based Activity Recognition System for Improving Health and Well-Being. 682–688. https://doi.org/10.1109/DSAA.2016.89

[57] Kelly White, Samuel Tate, Ros Zafonte, Shrikanth Narayanan, Matthias Mehl, Min Shin, and Amar Dhand. 2023. SocialBit: protocol for a prospective observational study to validate a wearable social sensor for stroke survivors with diverse neurological abilities. *BMJ Open* 13, 8. https://doi.org/10.1136/bmjopen-2023-076297

[58] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias (Eds.). European Language Resources Association (ELRA), Genoa, Italy, 1556–1559. http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf

[59] Chenren Xu, Sugang Li, Gang Liu, Yanyong Zhang, Emiliano Miluzzo, Yih-Farn Chen, Jun Li, and Bernhard Firner. 2013. Crowd++: unsupervised speaker count with smartphones. In *2013 UbiComp* (Zurich, Switzerland) *(UbiComp '13)*. ACM, New York, NY, USA, 43–52. https://doi.org/10.1145/2493432.2493435

[60] Ruiqing Yin, Hervé Bredin, and Claude Barras. 2017. Speaker Change Detection in Broadcast TV using Bidirectional Long Short-Term Memory Networks. In *2017 INTERSPEECH*. Stockholm, Sweden. https://github.com/yinruiqing/change_detection

[61] Huanle Zhang, Wan Du, Pengfei Zhou, Mo Li, and Prasant Mohapatra. 2016. DopEnc: Acoustic-Based Encounter Profiling Using Smartphones. In *2016 MobiCom* (New York City, New York) *(MobiCom '16)*. ACM, New York, NY, USA, 294–307. https://doi.org/10.1145/2973750.2973775