Automated Face-To-Face Conversation Detection on a Commodity Smartwatch with Acoustic Sensing

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

Understanding social interactions is relevant across many domains and applications, including psychology, behavioral sciences, human computer interaction, and healthcare. In this paper, we present a practical approach for automatically detecting face-to-face conversations by leveraging the acoustic sensing capabilities of an off-the-shelf, unmodified smartwatch. Our proposed framework incorporates feature representations extracted from different neural network setups and shows the benefits of feature fusion. The framework does not require an acoustic model specifically trained to the speech of the individual wearing the watch or of those nearby. We evaluate our framework with 39 participants in 18 homes in a semi-naturalistic study and with four participants in free living, obtaining an F1 score of 83.2% and 83.3% respectively for detecting user's conversations with the watch. Additionally, we study the real-time capability of our framework by deploying a system on an actual smartwatch and discuss several strategies to improve its practicality in real life. To support further work in this area by the research community, we also release our annotated dataset of conversations.

CCS Concepts: • **Computing methodologies** → *Neural networks*; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: conversation detection, audio sensing, speech, social interactions

ACM Reference Format:
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 (September 2023), 29 pages. https://doi.org/10.1145/3610882

1 INTRODUCTION

Social interactions are communication exchanges between two or more individuals and play a critical role in society [9]. They allow members of a community to socialize and provide a way for the spread and strengthening of cultural norms, values and information. Because of its importance, the ability to passively and objectively track and quantify face-to-face social interactions would be a breakthrough across several disciplines including behavioral sciences [14, 26], information propagation and diffusion [67, 76], social network analysis [11, 73], and the study of health and well-being [63, 69].

One of the fundamental components of social interactions is interpersonal communication, particularly face-to-face spoken communication [65]. While face-to-face conversations have been traditionally studied and documented via self-reports [14, 59], these methods pose a high burden on individuals and introduce biases in the data. Recent approaches have leveraged mobile devices to passively capture speech or social patterns *in situ* [7, 53, 54, 57, 59]. However, these methods suffer from shortcomings when it comes to inferring face-to-face conversational events in real life settings. Firstly, automated speech detection is not sufficient to characterize a

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

2474-9567/2023/9-ART109
https://doi.org/10.1145/3610882

conversation, since the presence of speech alone is not indicative of an interaction. Secondly, although researchers have proposed methods to gather information about social interactions with smartphones and collaborative sensing devices [8, 12, 25, 28], none of the prior efforts have investigated the feasibility of leveraging commercial on-body sensing platforms such as smartwatches, which can be less obtrusive for longitudinal deployment. Correspondingly, very little research has focused on modeling conversations with lightweight wearable devices, which poses additional constraints in terms of engineering design optimization. Lastly, the naturalistic setting of social participants in their respective environments makes conversation detection more challenging than traditional speaker diarization setups [4, 48].

To address the above research problems, we explore a practical approach for automatically detecting face-to-face conversations by leveraging the acoustic sensing capabilities of an off-the-shelf smartwatch. In recent years, smartwatches have been adopted in larger numbers and become increasingly more sophisticated. As a result, thanks to improvements in sensing and computational capabilities, these devices have become a very compelling platform for the real-time monitoring and inference of a myriad of human behaviors. In this work, we present a customized model that shows the benefits of feature fusion for conversation detection on smartwatches. We also conduct analysis with smartwatch data in the real world and study various techniques to improve the practicality of our system. The contributions of our work are summarized below:

- A practical approach for detecting face-to-face conversations by leveraging acoustic sensing with an off-the-shelf smartwatch. Our approach relies on a customized feature fusion framework based on lightweight neural networks. We optimize deployment of this framework by exploring techniques for power usage optimization based on adaptive sampling and contextual cues.
- An evaluation of our framework through a semi-naturalistic study of 39 participants (18 groups) in their homes and an unconstrained free-living study of four participants, including one participant over a week. In the semi-naturalistic study, our framework achieves an 83.2% F1 score for detecting user conversations. In the free-living study, it achieves an 83.3% F1 score. To the best of our knowledge, this is the first in-depth study to evaluate acoustic model performance for real-life face-to-face conversation detection based on a smartwatch.
- A system implementation of our method that can run on a watch in real-time. With this implementation, we demonstrate several strategies to improve the system's practicality, demonstrating its ability to detect conversations with low latency and optimized power consumption.

2 RELATED WORK

Over the last two decades, numerous studies have demonstrated the power of smartphones, wearable devices, and personal computers to capture everyday human behaviors, activities of daily living, emotions, health states and more [10, 13, 37, 68]. Additionally, sensor data from these devices has also been shown to be predictors of contextual information about people and their surroundings [36, 50]. Despite these advances, the detection and quantification of social interactions, a core element of society and human life, remains largely elusive. In this section, we highlight relevant prior work related to this topic and compare it against the specific contributions of our acoustic-based method.

2.1 Non-Acoustic-Based Methods for Detecting Social Interactions

Kim *et al.* shows that motion signals can be used for the inference of collaborations [30]. Bio signals collected from the human body such as respiration and electrocardiogram (ECG) can also be used as indicators of social behaviors, including speech [7, 16, 22, 61]. However, a major downside of such methods is that a dedicated device is required to be attached to the body for continuous collection of physiological inputs. Researchers have also explored the usage of radio signals captured by personal devices for monitoring people's interaction behaviors.

Radio signals are shown to be particularly useful for monitoring interactions in the same physical location [25, 40, 47, 66]. Besides, studies by Palaghias *et al.* and Yan *et al.* validated that Bluetooth on different devices can also be used to detect close contact among human subjects [58, 75]. However, the fact that radio and Bluetooth systems typically rely on a collaborative sensing framework with multiple devices can be an obstacle to scalability. More importantly, such modalities are often more related to the physical context rather than the actual spoken components of interactions.

2.2 Social Context Detection with Acoustic Sensing

The usage of audio for inference of everyday human physical activities and contexts has been widely studied [2, 36, 42, 45, 50]. To record data for longitudinal studies, Mehl *et al.* [54] developed the Electronically Activated Recorder (EAR). It is a portable chest-worn recorder that can capture user audio once every 12 minutes for two to four days. Similarly, Feng *et al.* proposed a system for continuous audio recording on a resource-constrained mobile device [17]. The authors specifically developed a strategy to reduce the power consumption of the device by only triggering the feature calculation steps during the target acoustic events. By using neural networks, Lane *et al.* deployed a system for acoustic sensing on a smartphone [35]. Their system was tested with a variety of acoustic sensing tasks, including classification of some common acoustic events, emotion recognition, and speaker identification.

Audio can also be used to reveal specific human behaviors related to the social interaction process. For example, Ahmed *et al.* [4] and Rachuri *et al.* [60] developed systems combining audio and sensor signals to recognize one's emotional states. Lu *et al.* [49] studied the viability of detecting stress from social interactions. Based on multiple mobile platforms, Lee *et al.* [39] proposed *Sociophone*, a sensor fusion framework to detect users' meta-linguistic contexts including their turn-taking behaviors and roles. A similar effort was proposed by Li *et al.* [41], but the authors applied multiple sensor boards attached to human bodies to capture the signals. Xu *et al.* used audio collected on a smartphone to estimate the number of speakers in a meeting [74]. Hsiao *et al.* [24] showed that it is possible to detect the social engagement level of a group conversation. They were able to extract turn-taking features for a conversation group based on audio data collected by a smartphone. By training on user samples, speaker identification techniques can also be adopted on real-time systems to infer the user identity [4, 48, 52].

Our work differs from the above efforts in several respects. Firstly, unlike the existing efforts [4, 24, 49, 60, 74] which focus on social contexts or user emotional factors, we focus on the detection of physical user communications (e.g., through monologues or two-way face-to-face conversations). Secondly, unlike the above speaker identification systems [4, 48, 52], we do not aim to build a speaker classifier to identify individuals in the interaction process, which requires preliminary information collected from the speakers. In other words, our work characterizes user behaviors in a purely speaker-agnostic phase. Thirdly, our study is based on a single device rather than a collaborative framework of multiple devices, as in [39, 41]. Fourthly, our framework is based on a smartwatch, which is not explored by any of the above efforts.

2.3 Voice Activity Detection

Voice activity detection is a research area that has been extensively explored. Today, the human voice can often be recognized within a sound classification task by a general-purpose audio sensing framework [50, 51]. Sehgal *et al.* [64] publicized a well-trained voice activity detection (VAD) app for commercial smartphones. However, the general VAD process falls short of identifying the source of the voice. For example, such existing solutions cannot identify if the detected voice is from the user of a device or from someone else talking in the background. This makes it challenging to extend the existing work for modeling spoken communications, which is mostly characterized by the turn taking of different speakers [38]. Recently, Nadarajan *et al.* [56] studied the feasibility of detecting voice activities related to a user based on audio collected by a chest-worn badge. Similarly, Little *et al.*

[46] proposed an approach based on a custom-designed wrist-worn device, and they showed that it is possible to identify a user's speech from the background sounds by training on data collected from a large group of wearer and non-wearer participants.

Unlike all the previously mentioned efforts, our work aims to infer user behaviors related to *interactions* such as one-way speech and conversations, by making use of an off-the-shelf smartwatch. Besides, unlike the above studies that are developed for specific use cases such as meetings [56] or patients of depression [46], we examine our conversation detector based on varying unconstrained social events in everyday life. Moreover, we demonstrate the real-time capability of our system for a commercial smartwatch, which was not presented by any of the above work.

2.4 Face-To-Face Conversation Detection with Acoustic Sensing

To recognize spoken interactions among individuals, a common method is aggregating information from multiple devices. For example, voice frame matching [72] and mutual information of audio signals [8, 12, 52] collected by a pair of audio recorders can be indicators of whether two individuals are involved in the same conversational session. Recent work has also explored the usage of vocal features [4, 39, 41], Doppler profiling [79], and the fusion of audio and Bluetooth signals [52] collected from multiple smartphones to infer interactions among individuals. Despite the promising results obtained by these existing solutions, the fact that multiple devices have to be triggered simultaneously can be an obstacle for scaling and personal usage. In contrast to these approaches, our work studies the potential of modeling spoken interactions between a user and other individuals based on acoustic data collected by a single smartwatch. Recent commercial products such as the Apple AirPods are announced to be equipped with conversationally-aware features for better user experience [5]. To the best of our knowledge, however, our study is the first in the literature to quantify the performance of face-to-face conversation detection in unconstrained daily living scenarios by using audio data collected from a watch.

3 AUTOMATED CONVERSATION DETECTION

3.1 Task Formulation

In daily living scenarios, not all speech characterizes conversations, and not all speech captured by a smart device originates from the user wearing the device. Admittedly, the definition of conversation can vary depending on context [70], but a common agreement in the literature is that a conversation should be a spoken interaction between at least two participants [15, 70, 71] with valid turn-takings between the speakers [15, 21]. By considering common scenarios in practice, the detection model of our system is developed to be a three-class classifier. In our study, we define *conversation* as our first target class, where audio instances of this type should contain in-person spoken communications with valid turn-takings between the device user and at least one other participant. Specifically, an instance without any speech turns taken by the user (e.g., conversations recorded on a television) is *not* considered to be a valid conversational instance in our study, since we emphasize the detection of conversations that involve the user. In addition to conversations, we refer to our second target class as *other speech*. This includes speech instances generated by the device user that do not contain valid speech turns by other social participants. Common examples in our study include user activities of reading or story-telling. Other recorded sound types are categorized as *ambient sound* in our study, including ambient noise, silence, or background voice such as TV sounds or someone talking nearby.

Our system is developed such that a user does not need to adapt the system to his or her own voice or the acoustic environments. Besides, we focus on *instance-level* recognition, where we apply a fixed granularity of 30 seconds as the recognition window. Selection of the window size is empirical, and a window size of 30 seconds has been commonly used in prior work on conversational analysis [7, 12, 61] since it provides a reasonable balance between the precision of inference and the coverage of turn-takings in a conversational episode.

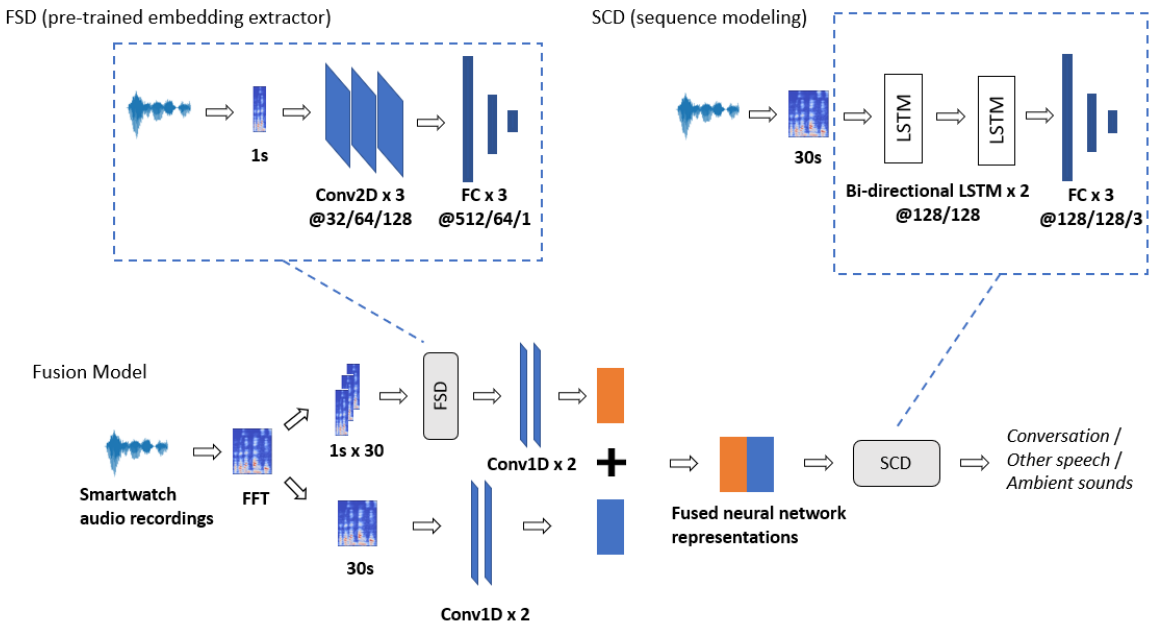

Fig. 1. Overview of our proposed framework (the fusion model). It is based on a customized neural network architecture that fuses feature representations extracted from different model setups. Conv: convolutional layers; FC: fully-connected layers; FSD: foreground speech detector; SCD: speaker change detector.

3.2 Conversation Modeling

As discussed earlier, the canonical approach of conversation detection with acoustic sensing is comparing the correlation of voice streams recorded simultaneously by multiple social participants to determine if someone is in a conversational event with others. However, this was often conducted in a controlled environment and required dedicated recording devices for individual participants. By including modern deep learning techniques, we re-consider this problem for a more realistic setting based on a single device. Specifically, three research questions are explored:

(1) What is the performance of conversation detection in the wild by applying existing deep learning models trained exclusively on public acoustic event datasets?
(2) By using neural networks, can we obtain a reliable inference performance for conversational data while maintaining a reasonably compact model size for a smartwatch?
(3) Is it possible to capture additional speech features such as speech proximity to the smartwatch and apply feature fusion to enhance conversation detection for the user of the watch?

To explore the first question, we studied two popular deep learning models, *CNN14* [33] and *YAMNet* [19], which are both developed based on the large-scale public YouTube AudioSet [18]. In addition, we leveraged a pre-trained YAMNet model as a feature extractor and built a customized neural network classifier based on our user dataset for a comprehensive comparison. Details of the model architectures and setups will be presented in Section 3.3.

For question 2), we compared the usage of a convolutional neural network (CNN) and sequence models for conversation modeling. The CNN we tested was based on the original *MobileNetV1* [23], a general-purpose deep

learning model architecture optimized for mobile platforms. Different from the pre-trained YAMNet, this model was built with our user dataset from scratch. Sequence models such as the long short-term memory (LSTM) have been demonstrated to be effective for the modeling of speech turns [77]. This also inspires our study. Since in-person human conversations are formulated and can be characterized by the transition of speech turns [15, 21], a model dedicated to speech turn modeling may be a good candidate for conversation detection as well. We thus adopted the model architecture of a **speaker change detector** (**SCD**) [77], originally used to detect speaker change points in audio frames, as one of our baselines. A speaker change point is the boundary of speech turns for different speakers in a conversation. Correspondingly, speaker change detection is a binary classification task aiming to infer the existence of such boundaries in audio frames.

The exploration of question 3) is based on prior work [56] that presents a method of detecting user's speech on a wearable device by distinguishing voice activities in close proximity to the device (typically the wearer's voice) from the background sounds. The detection of the user's voice activities is thus independent of the user's voice fingerprints, and the model trained to classify the foreground wearer's voice and background sounds is referred to as a **foreground speech detector** (**FSD**). Although the original study was aimed at chest-worn recorders in meeting environments, we believe that this same method may also be applied to smartwatches and to obtain indicators of whether the device wearer is involved in a conversational event or not. By adopting the FSD, we can capture the wearer's speech cues on a watch without building a model with the user voice samples *apriori*.

Following the setup of the FSD, we further investigated a strategy for feature fusion based on a customized neural network architecture (Figure 1). The customized fusion model consists of a sequence architecture, and is jointly optimized with general-purpose acoustic features and knowledge representations obtained by a pre-trained FSD. In the next section, we will describe the detailed architectures and setups for each of the models.

3.3 Detailed Model Architectures

3.3.1 Models Trained on AudioSet. The Google AudioSet is one of the biggest public acoustic event datasets nowadays, containing over 2M 10-second clips of soundtracks extracted from public YouTube videos. The dataset is highly class-imbalanced, including a total of 527 human-labeled sound classes. Several sound classes are relevant to human speech and conversations, such as *Speech, Conversation, Shout, Narration*, etc. Given the rich acoustic context contained in AudioSet, we first explored implementing conversation detection models trained with this dataset only. In our study, we chose CNN14 and YAMNet. CNN14 is a general-purpose deep acoustic event detector developed and evaluated based on AudioSet, achieving state-of-the-art detection performance. It is a stack of multiple convolutional layers, with around 80M trainable parameters. YAMNet, on the contrary, is based on the MobileNetV1 CNN architecture for mobile devices. Although the model is also trained on AudioSet, it is much more lightweight with around 3M trainable parameters.

We used the public source code of both models[1][2] to set them up for inference. For both models, the input audio was sampled at 16 KHz mono. The model inputs were spectrogram features extracted from the audio. To handle the class label mismatch between AudioSet and our study, we categorized the AudioSet classes into three types. Specifically, an inferred class of either *Conversation, Chatter*, or *Whispering* was considered equivalent to our target class of *conversation*. The AudioSet classes of *Speech, Shout, Screaming, Laughter, Narration / monologue*, and *Crying / sobbing* were equivalent to *other speech*. All other AudioSet classes were categorized as *ambient sounds*. Since our target classes are mutually exclusive, we counted the top-1 class predictions from the models as the final inference results.

3.3.2 Foreground Speech Detector (FSD) Baseline. The original FSD takes as inputs a one-second audio instance each time and generates an inferred probability of foreground speech. In our study, we modified the shape of the

[1] https://github.com/qiuqiangkong/audioset_tagging_cnn
[2] https://www.tensorflow.org/hub/tutorials/yamnet

neural network layers to fit for our 30-second conversational instances. Additionally, while the original setting of the model was limited to certain statistical and spectral features due to privacy constraints, we used more descriptive spectrogram features in our study.

Our FSD consisted of three 2D convolutional layers with the ReLU [3] activation (Figure 1, top left). The kernel size for each layer was (4×4), with a stride of (2×2). The outputs of the layers were padded to be of the same size as the inputs. Besides, each convolutional layer came with batch normalization and a max-pooling operation of (2×2) kernel with the same stride. The fully-connected layers were activated by the ReLU activation except for the output. As a baseline model for conversation detection, the FSD had three neurons as output. As a feature extractor, however, the FSD had only one neuron as output. The fully-connected layers were also followed by batch normalization. To connect the convolutional layers and the fully-connected layers, the outputs of the last convolutional layer were flattened. The model had 0.5M trainable parameters.

3.3.3 Speaker Change Detector (SCD) Baseline. The SCD consisted of two bi-directional LSTM layers and three fully-connected layers (Figure 1, top right). Different from the FSD, the extra LSTM layers enable the SCD to better capture the speaker turn patterns in an audio sequence. The output sequence of the LSTM layers was directly passed to the first fully-connected layer without flattening. Outputs of the first two fully-connected layers were activated by the Tanh activation function, and then globally averaged along the temporal dimension. For the baseline test in our study, the output layer was modified as three neurons, activated by Softmax. The number of trainable parameters was 0.7M.

Inputs to both the FSD and the SCD were the same spectrogram features extracted every 30 seconds. We sampled the input audio at a sampling rate of 16 kHz, and then framed the audio at a size of 8,000 samples with 50% hop for the fast Fourier transform (FFT). The number of frequency bins was kept as 128, resulting in image-like spectrogram features of shape (128×120) per instance. Before fitting a model, we normalized the FFT features globally with their mean and standard deviation.

3.3.4 The Fusion Model. As mentioned earlier, we explored model fusion strategies based on the original FSD and SCD to augment the overall conversation modeling performance. Canonically, feature fusion can be implemented at different stages of classification, from the input stage to the output decision stage [10, 29, 43]. In our study, we compared feature fusion at the intermediate layers of the neural networks (our proposed architecture) and at the final output layers of the individual models, respectively. Our study showed that the intermediate fusion yielded the best inference performance with a smaller model size (results in Section 5). To enable the fusion, we designed a customized neural network architecture (Figure 1, bottom), maintaining a similar model size compared to the individual baselines. The new fusion model consisted of two branches, each responsible for the foreground representations obtained by the FSD and the general-purpose acoustic spectrogram, respectively. Specifically, the FSD in the fusion model was developed with a separate foreground dataset we prepared ahead of time. While training the fusion model, the FSD was fixed and only used as a knowledge extractor. The output representations of each branch were then concatenated along the temporal dimension and fed to a stack of LSTM layers.

Both branches of the fusion model fit with the same spectrogram features as mentioned above, but we sliced the spectrogram to be 30 1-second clips for foreground knowledge extraction to improve the temporal precision of the representations. Hence, the input shapes per instance were (128×4) and (128×120), respectively, for the two branches of the fusion model. In our study, we extracted embedding features from the first fully-connected layer of the FSD instead of the last, since these yielded a more stable training performance for the fusion model. Before concatenation, the extracted foreground representations were stacked every 30 seconds to match the output shape of the other branch. The 1D convolution was performed along the temporal dimension of the features in each branch. The number of trainable parameters was 0.8M for the fusion model, which is lightweight for real-time deployment on edge devices.

Table 1. Scripted activities and the corresponding time schedule of the activities for our audio collection.

Schedule	Scripted Activities
3:30pm	Reading out loud a text passage
4pm	Making a telephone call
4:30pm	Watching TV
5pm	Face-to-Face Conversations
5:30pm	Outdoor Conversations
6pm - 8pm	Meal Preparation and Dining

3.3.5 Other Models Developed with Our User Dataset. In addition to the mentioned models, we developed and evaluated two extra models based on our user dataset. The first was MobileNetV1, which we trained from scratch. The second was YAMNet trained on AudioSet as an acoustic feature extractor, followed by a customized fully-connected neural network classifier developed on our dataset. Specifically, we extracted clip-averaged acoustic embedding features of 1,024 dimensions from YAMNet. The classifier consisted of three fully-connected layers of size 128, 128, and 3, respectively. Inputs to both models were the same 30-second spectrogram, as described.

4 SEMI-NATURALISTIC DATA COLLECTION

4.1 Study Script

Following the design of our models, we then conducted a semi-naturalistic study for a rigorous and comprehensive evaluation of their performance. This IRB-approved study took place in the homes of 18 groups of participants. Each group consisted of one primary individual wearing a commercial smartwatch (Google Fossil Watch Gen 4) with audio capture capabilities, i.e., the device user, and at least one more individual in the home. Participants were asked to follow a script of six activities, shown in Table 1 and detailed below.

(1) **Read a Text Passage**: The user of the watch was required to read aloud a 434-word excerpt from a Wikipedia page, which typically lasted between two and three minutes.
(2) **Place a Telephone Call**: The second scripted activity was to place a telephone call between the smartwatch user and another remote researcher. The script did not specify any requirements regarding the location or content of the conversation, allowing the call to be performed in a completely natural way.
(3) **Watch TV**: The third activity entailed watching content on a TV or laptop with the sound on for around 10 minutes. Conversations were allowed during the activity.
(4) **Conversation Indoors**: In the fourth activity, indoor face-to-face conversation, all participants of the group played the NASA decision-making survival game [20] while sitting around a table. This activity was designed to shed light on the speech patterns of natural interactions within a group of people, so all participants were encouraged to talk during the discussion. The motivation of this setup was to collect people's speech patterns while naturally interacting with others.
(5) **Conversation Outdoors**: For the fifth activity, participants were instructed to walk outdoors for at least five minutes and chat with each other on any topic of their choosing.
(6) **Meal Preparation and Dining**: The final scripted activity involved meal preparation and dining, where the user was instructed to take charge of cooking for the night of the study. This activity enabled us to collect conversational audio mixed in with kitchen and cooking sounds.

Although the activities were pre-specified, we emphasized the naturalistic aspect of data collection. The activities took place in participants' own home environments and were performed without the presence of the research team. Hence, other household members were free to join and talk to the specified study participants without restrictions. Additionally, the scripted activities were timed to align with people's actual daily routines, and we left 30-minute gaps between each scripted activity to ensure that participants had sufficient time to complete them. For all of the activities except TV watching, the script did not specify a set time for the activities to end.

4.2 Protocol and Procedures

All of our studies were conducted remotely. Before each study, a researcher delivered the smartwatch, the text materials for scripted activities, and the study instructions to the participants. On the day of the study, the researcher connected with the participants remotely via video-conference. The researcher then described the details of the study, answered any questions from the study participants, and taught the users how to complete the smartwatch setup, a process that typically took 10-15 minutes. Then, the researcher logged out of the meeting and the recording began.

To facilitate data collection and make the study as naturalistic as possible, audio was recorded with a smartwatch continuously throughout the day. For activities that demanded privacy (e.g., going to the bathroom), the users could temporarily remove the watch and/or turn it off. Also, we required that the watch not be covered by clothing to eliminate noises produced by friction with clothing and to ensure that the activity sounds would not be muffled. The users were asked to check the status of the app before each specified activity to ensure that the app was recording properly by checking to see if the timer on the app was running. To remind the users to check the app and to simplify our data annotation, we also asked them to roughly write down the time stamps of the timer when they checked it. In the event of technical issues or questions, participants were able to reach researchers via text or phone call. After 8pm, the users stopped the recording and the study was complete. The data was saved automatically after the app was stopped. A researcher picked up the study equipment and materials from participants after the completion of the study and had them sanitized and prepared for the next study.

4.3 Participants

In total, we collected data from 18 groups of participants. Overall, we did not pose any requirements on their age, handedness, occupation, or number of household members in their homes. By working together with a local recruiting agency, we enrolled a diverse set of participants including college students, an engineer, a book keeper, a travel agent, a local government campaign manager and a tattoo parlor owner. Participants' ages ranged from 15 to 59, and we aimed for gender balance in our study cohort. All study participants were fluent in English or native English speakers. Table 2 provides additional details about the study groups. 10 out of the 18 groups consisted of only two study participants, while the remaining contained three or more social participants, including extra household members throughout the study process.

4.4 Data Annotation

After each group of data collection, the audio was transferred to a server from the smartwatch, where we first determined the audio segments of the scripted activities by listening back to the full audio clips. The time points noted down by the participants were also used as reference points for better segmentation of the audio. In the next step, we examined each of the audio segments and annotated the audio instances. Our dataset was used for two purposes- first to build the FSD and then to evaluate our conversation models. Hence, two groups of researchers were engaged to annotate the data, each including two independent annotators. The first group

Table 2. Overview of the collected study participants (Participants: number of participants recruited for the study; Household: number of actual household members of a group; M: Male, F: Female). We required diversity in gender and age to ensure better generalization of our collected data on the public.

Group #	Participants	Household	Gender	Age	Handiness
1	2	2	F, F	18, 20	Right
2	2	2	F, F	19, 20	Right
3	2	6 (1 baby)	F, F	37, 14	Right
4	2	3 (1 baby)	F, M	29, 32	Left
5	2	3	M, M	36, 15	Left
6	2	4	M, F	41, 15	Right
7	2	2	F, M	55, 18	Right
8	3	3	F, M, F	45, 43, 10	Left
9	3	3	F, M, M	55, 59, NA	Right
10	3	5 (1 baby)	M, F, M	44, 25, NA	Right
11	2	2	M, F	41, 41	Right
12	2	2	F, F	65, 64	Right
13	2	2	M, F	26, 26	Right
14	2	2	F, M	19, 28	Right
15	2	2	M, F	21, 22	Right
16	2	2	M, M	24, 22	Right
17	2	2	M, M	26, 24	Right
18	2	3	F, M	26, 26	Right

annotated the audio data at a granularity of one second to determine instances of foreground user speech. The second group annotated the audio at a granularity of 30 seconds for the three target classes of our study.

4.4.1 Annotation of Foreground Speech. At each second, our annotators assigned one of the three labels *foreground speech*, *other sounds*, and *ambiguous sounds* to the clips by listening to the audio content. *Foreground speech* includes situations when only the device user talked within the second, or if there was an overlap between the user and other human speakers. All other sound types, including speech from other participants and ambient noise, were categorized as *other sounds*. Instances of silence were also counted as *other sounds*. The label *ambiguous sounds* was used for instances where the annotators were not confident about the sound type. This was observed for some study groups when the voice between the smartwatch user and the other participants was too similar to each other, since the study was conducted remotely and we could not monitor the entire process.

To verify the quality of annotating the foreground instances, we selected one interaction session from a participant group and computed the inter-rater reliability between the annotators. We used the Cohen's kappa to measure the pair-wise annotation quality of our data. Compared to some other measures such as the joint probability of agreement, the Cohen's kappa can be more robust to the effects of agreement by chance, especially given our small number of target classes. We obtained a mean of 0.91 Cohen's kappa, indicating good agreement amongst the annotators [34].

4.4.2 Annotation of Face-to-face Conversations. For every 30 seconds, our annotators applied three mutually-exclusive labels: *conversation*, *other speech*, and *ambient sounds*, as defined earlier. Again, labels *conversation* and *other speech* were assigned with an emphasis on speech truly generated by the user. Speech or conversations

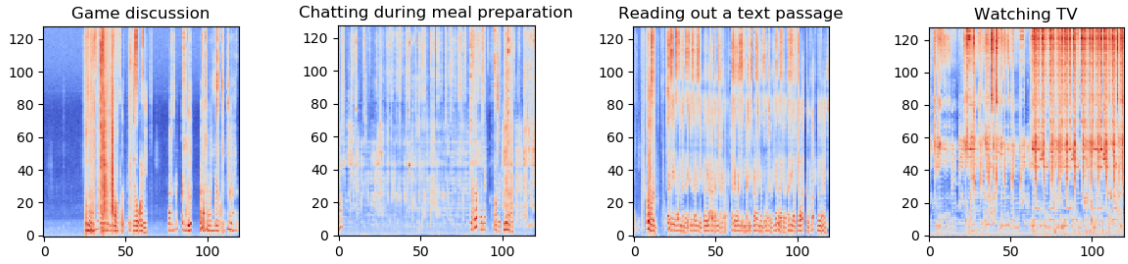

Fig. 2. Sample spectrogram features used by our models, extracted from different instance types. All features are converted to the log scale for better visualization. Y-axis: Frequency bins; X-axis: Frames.

captured on the television or from irrelevant participants in the house were included in *ambient sounds*. Additionally, *ambient sounds* also included noise of appliances in the house, activities of daily living sounds, and white noise. In general, we did not count single words or ritualised speech such as 'hey', 'emm', 'ahh', and laughing or coughing sounds as a valid speech turn. In rare cases, we noticed that there were simultaneous conversational streams of the users and other speech types in the same 30-second window. Such clips were still annotated as *conversation* given that a conversational event was included. It is noted that our annotation was not bounded by the types of scripted activities. Instead, it was only based on the speech turn patterns of the instances.

To verify the quality of the annotated conversational instances, we randomly selected samples from two interaction groups to compute the inter-rater reliability. We observed a similarly high agreement level of 0.88 Cohen's kappa.

4.5 Collected Data

In total, we obtained around 32 hours (114,637 seconds) of audio recordings for the semi-naturalistic study, including 26,982 user foreground speech instances. Additionally, we obtained 1,780 instances of *conversation*, accounting for 45.7% of the total audio. *Other speech* and *ambient sound* accounted for 19% and 35%, respectively. Figure 2 shows sample spectrogram features used by our models, extracted from different instance types[3]. The features were converted to the log scale for better visualization. We notice that the instance of *other speech* (e.g., user reading out a text passage) tends to have a more uniform frequency distribution within the vocal region over time than those of conversational instances (e.g., game discussion and chatting during meal preparation). This indicates that the spectrogram features may capture the transition of speaker turns in conversations, which is reflected in the inconsistency of the frequency distributions. In addition, we observed that instances of *ambient sounds* can be distinct from *conversation* and *other speech*. An example is watching TV, where the vocal frequency distribution of the television is highly distinct from those of the remaining three instances.

5 PERFORMANCE AND ANALYSIS

In this section, we describe the experiments we conducted for conversation detection with our semi-naturalistic data. Specifically, we evaluated the performance of our model and compared it against the other baselines by aggregating all the participants' data, since this gives us a better sense of how the models perform over the study participants in general.

[3]More samples can be accessed at: https://doi.org/10.18738/T8/U6CFIP

5.1 Evaluation Setup

As described earlier, we built a speech representation extractor based on the FSD architecture for our fusion model, so the training of the representation extractor and the conversation models was separate. For all conversation models developed on our user dataset, we applied the categorical cross-entropy loss with an Adam [31] optimizer. The learning rate we used was 0.01, with the Beta values of (0.9, 0.999). We trained a model until it reached either a maximum epoch number of 100 or a best validation accuracy for 15 consecutive epochs. The batch size was 128. The training setup for the representation extractor was similar, except that we used binary cross-entropy as the loss for foreground detection and switched its output activation to Sigmoid. To create the models, we used the TensorFlow [1] Keras API in Python.

To evaluate the conversation models, we followed a leave-one-group-out (LOGO) evaluation scheme, where all but one group of study participants were used for model training and the remaining group was used to derive checkpoint models. In each LOGO fold, we reported the macro (unweighted) average of class F1 scores, macro class recall, and macro class precision as the performance metrics for the multi-class setting. Using the macro average enables us to understand the model performance by treating all target classes with equal importance, eliminating the effects of imbalanced class distributions of the evaluation set. We calculated the mean of the metrics for all groups of participants and then reported the global mean for the LOGO study.

By training with the foreground dataset, the foreground representation extractor we used for our fusion model achieved a macro F1 score of 82.0% for the binary foreground detection over all participants. The result demonstrates the effectiveness of the model in capturing the foreground speech cues based on the smartwatch audio recordings. Hence, we leveraged its extracted representations to enhance the fusion model for conversation detection.

5.2 Overall Model Performance

The LOGO experiments were conducted individually for each conversation model, and Table 3 shows the results of the experiments. For CNN14 and YAMNet trained on AudioSet, only inference was performed. To reiterate, the three research questions we hope to address are:

(1) What is the performance of conversation detection in the wild by applying existing deep learning models trained exclusively on public acoustic event datasets?
(2) By using neural networks, can we obtain a reliable inference performance for conversational data while maintaining a reasonably compact model size for a smartwatch?
(3) Is it possible to capture additional speech features such as speech proximity to the smartwatch and apply feature fusion to enhance conversation detection for the user of the watch?

Regarding research question 1), we can see from Table 3 that the classification results obtained by CNN14 and YAMNet trained on AudioSet are the worst among all tested models. The higher model complexity of CNN14 does not improve its performance for recognizing the user's conversations. As a comparison, the combination of YAMNet as an acoustic feature extractor and customized fully-connected neural network significantly improves the recognition performance from an F1 score of 28.9% to 65.5%, with a comparable model size (3.3M to 3.4M). The results and comparison indicate the significant domain shift between conversational data obtained from public YouTube videos and conversations from audio recordings collected by a smartwatch in daily life. Customization of the models based on real-life user data can be very helpful for a more reliable model performance.

To explore research question 2), we compared the model performance based on our user dataset. It can be seen that the best recognition performance is obtained by the two sequence models, SCD with LSTM and the fusion model. A possible explanation is that the temporal patterns of the audio streams may reflect the change of speech turns which can be used to characterize human conversations, and these patterns may be better captured by the sequence models. Furthermore, the sequence models obtain a reasonable inference performance with a relatively

Table 3. Overall classification performance for *conversation*, *other speech*, and *ambient sound*; data based on the 18 groups of study participants. The fusion model achieves the best inference performance with a compact model size, which is preferable for deployment on wearable platforms.

Model	Macro F1	Macro Recall	Macro Precision	# of Parameters
CNN14 (AudioSet)	23.9%	42.4%	30.3%	79.7M
YAMNet (AudioSet)	28.9%	44.6%	28.4%	3.3M
FSD	61.7%	63.1%	64.9%	0.5M
YAMNet (AudioSet) + FC	65.5%	66.2%	76.4%	3.4M
MobileNetV1	66.3%	67.7%	68.5%	3.3M
SCD (LSTM)	71.6%	71.3%	77.6%	0.7M
The Fusion Model (Proposed)	**76.2**%	**77.3**%	**77.7**%	0.8M

compact model size, especially compared to the tested general-purpose deep CNN. This is more desirable for on-device deployment.

Compared to the other models, our fusion model achieves the best inference performance for the LOGO study, with an F1 score of 76.2%, a macro class recall of 77.3%, and a macro class precision of 77.7%, for all participants. The better results show the effectiveness of model fusion over the individual baseline models. Specifically, the inclusion of the foreground features provides additional user speech cues, which are effective for the model to determine user involvement in a conversational event.

Last but not least, the model size is also an important factor to consider for model deployment on real devices. Table 3 summarizes the number of trainable parameters for each model. In general, the fusion model achieves good inference performance with a reasonable model size. As mentioned earlier, we also conducted an experiment based on decision-level fusion of the FSD and the SCD. To do so, each of the baselines was connected with individual LSTM and fully-connected layers, and the output class probabilities were averaged between the two. The macro F1, recall, and precision of the decision-fusion model were 75.2%, 75.8%, and 76.8%, respectively, for the LOGO experiment. The model size was 1.3M. Hence, by comparing the inference performance and the model size, the intermediate fusion approach is a better choice for our conversation detector. The next section will be focused on detailed analysis of the model performance for the semi-naturalistic study.

5.3 Result Analysis

To better understand the performance of our conversation detector, we examined the class-wise performance by aggregating results over all groups of study participants (Figure 3). From the confusion matrices, we can see that the model performance for classes *conversation* and *ambient sounds* is the best, with a class recall of 85% and 88%, respectively. Recognizing instances of *other speech* is the most challenging. By listening back to the raw data, we found that several situations of *other speech* in the study could be confused with *conversation*. For example, there were situations such as a user giving a telephone call (with no voice captured by the watch from the other speaker), a user commenting on a TV series (with no response from other participants), or a user giving a one-way instruction to others without seeking feedback. Although these instances were counted as non-conversational types, their similarity with the conversational events could affect the model inference. The right plot of Figure 3 shows the performance of the model for detecting conversations as a binary detection task. The class F1 score and accuracy for detecting *conversation* are 83.2% and 85.0%, respectively, with a false positive rate of 15.0% and a false negative rate of 16.3%. Although not strictly comparable, a recent method of conversation detection following the same granularity (30 seconds) [7] achieves 87% detection accuracy, a result on par with

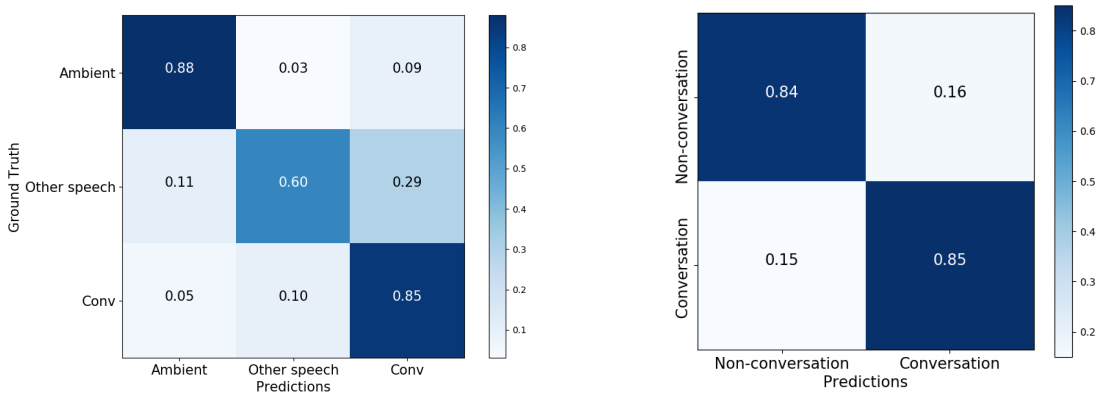

Fig. 3. Confusion matrices of inferring classes *conversation* (conv), *other speech*, and *ambient sound* (ambient) (left), and detecting conversations specifically (right). The data is based on our semi-naturalistic study with all 18 groups of participants.

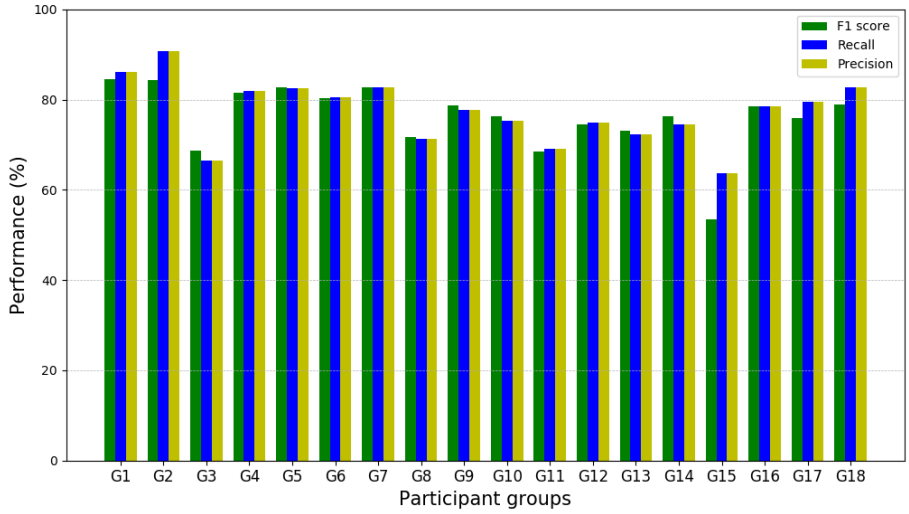

Fig. 4. Group-wise performance of recognizing *conversation* versus *other speech* versus *ambient sound* for the smartwatch users in the semi-naturalistic study.

ours. Their method used respiration signals collected from a wireless respiration sensor worn on the participant's chest, while we are focused on a practical system with acoustic sensing on a single commercial smartwatch.

Figure 4 plots the group-level inference performance for the semi-naturalistic study. The best performance is observed for group 2, with a macro F1 score of 84.4%. On the contrary, participant groups 3 and 15 are the worst, with a macro F1 score of 68.6% and 71.1%, respectively. To better understand the reason behind this, we

Table 4. Inference performance (3-class) of the fusion model architecture with further optimization; data based on the 18 groups of study participants. FC: fully-connected layers.

Fine-tuning Strategy	Macro F1	Macro Recall	Macro Precision
Quantization, mixed	73.7%	73.8%	76.8%
Quantization, FC only	75.2%	76.0%	77.2%
Weight pruning (selected)	**76.7%**	**76.8%**	**78.9%**

further listened back to the audio data. We found that the degradation of performance might be attributed to two main factors. Firstly, we found that the reliability of foreground detection could impact the overall performance of the conversation model. The F1 scores of foreground detection for groups 3 and 15, for example, were 68.6% and 53.5%, respectively, which were the lowest of all groups. The poor foreground results indicate that the user speech for these groups tends to be similar to the acoustic events in the background, making it challenging to recognize user conversations from the environments. Secondly, we noticed that the communication traits of different participants could vary significantly, making the inference more challenging. For example, while some participants were talkative and more detail-oriented during group discussions, some tended to be pure listeners, with little vocal responses in the sessions. Such sparsity of vocal responses led to highly imbalanced speaker turns in a conversational segment. In our study, we found that a conversation with sufficient turns back-and-forth was helpful for inference. This is reasonable, because it makes a conversation more distinct from other speech types such as a monologue.

5.4 Further Model Optimization

To facilitate model deployment on smartwatches, we further optimized our fusion model. Specifically, we deployed quantization-aware training [27] and weight pruning [80] to further reduce the model complexity without loss of inference performance. By applying quantization-aware training, we lowered the precision of parameters for the neural network layers. By applying model pruning, on the other hand, we optimized the model by zeroing out insignificant model weights, as this has been shown to be beneficial for the inference latency [80].

In our study, we followed the same training setup of our original fusion model for both quantization and weight pruning to obtain the new checkpoints. Checkpoints with both strategies were fine-tuned based on the pre-trained weights of our fusion model instead of training from scratch. We then conducted the same LOGO test with the semi-naturalistic user dataset. For quantization-aware training, we examined quantization on either multiple layers of the model or only the fully-connected layers. We converted the parameters from their original 32-bit float representations to 8-bit integer representations, including layer weights and activation. For pruning, we applied constant sparsity throughout training, where 50% of weights per layer were pruned. Furthermore, the pruning was applied for every 100 training steps. We used TensorFlow [1] to deploy the optimization.

Table 4 shows the change in the model inference performance with the two optimization strategies. Overall, our fusion model optimized with both quantization and weight pruning can still maintain the original inference performance. However, quantization for only the fully-connected layers is better than for mixed layers. This is possibly because the early layers of the fusion model are more critical in capturing the patterns of the input spectrogram and foreground features. Besides, we can see that the checkpoint model with weight pruning even slightly outperforms the original baseline. A possible explanation is that the reduction of model complexity may improve over-fitting and enable the model to focus on more important features. The benefits of the optimized checkpoints regarding actual model sizes and inference latency will be further discussed in Section 7.

Table 5. The majority of location contexts and activities of daily living captured in the free-living study.

Context	Major Activity	Percentage
Office	Working, Studying, Meeting, Vacuum cleaning	40%
Apartment	Watching TV, Having a meal, Chatting, Washing with water	23%
Street	Strolling, Chatting	8%
Bar/Restaurant	Having a party, Having a meal	8%
Vehicle	Driving, Listening to radio, Chatting	7%
Clinic (Dental)	Visiting a doctor	4%
Public building	Giving a presentation, Attending a workshop	4%
Grocery/Supermarket	Shopping	2%

6 FREE-LIVING STUDY

In this section, we describe our effort to evaluate our approach in real-world settings with four participants, including three participants with three hours of data each and one participant with 35 hours of data collected over a week. The goal of this extra study is to further examine how our model can be generalized in completely unconstrained scenarios, especially under conversational contexts that were not covered in the training pool of our semi-naturalistic study.

6.1 Data Collection

To collect data for the free-living study, four individual users wore the same Fossil smartwatch as in the semi-naturalistic study to record sounds in daily living. The users followed the same setup with the audio recording app for continuous and unobtrusive recording. For one participant (a researcher of the paper), the study lasted for a whole week, and the watch was worn during the daytime so that there could be a higher chance of capturing the user's social interactions with others during the data collection period. For the other participants, the study lasted for around three hours each. The participants were free to choose the time scheduled for the study, but they were required to choose a period when at least some conversations could be expected. All participants simply followed their daily routines to perform a variety of activities in daily living, including working, having meetings / discussions with other human subjects, studying, having meals with housemates, having a party, etc. For the one-week data collection, we did not specify an exact start and end time each day for recording, but most of the recording started before noon of a day and ended in the evening of that day to include the majority of contexts during the daytime. For the other participants, they were instructed to stop recording on the watch after roughly three hours of recording. Also, they were allowed to take off the watch for privacy reasons if necessary (e.g., in the bathroom) and resumed wearing the watch afterwards. To ensure the unconstrained nature of the study, we did not have further restrictions on the sensing environments or the social participants during the recording period. An IRB form was signed by all non-author participants.

In addition to the smartwatch, we also provided three participants with a Google Pixel 1st Gen smartphone equipped with our application. The goal of introducing this additional device was to understand how data collection and inference performance at conversation detection might differ across these devices. Once recording started on both devices, a marker sound was generated for synchronization. The participants were then instructed to keep the phone around similarly to how they would with their own personal smartphones. The phone recording was stopped whenever data collection with the watch was also completed. After the study, the participants were required to take a note of the way they placed and kept their smartphones. The phone study was covered by the

Table 6. Detailed inference performance by applying one of our checkpoint models (with pruning) to detect face-to-face user conversations versus non-conversations in the unconstrained free-living study, categorized by specific location contexts.

Context	Macro F1	Weighted F1	Macro Recall	Macro Precision
Overall	89.2%	92.4%	89.7%	88.7%
Overall (no pruning)	88.3%	91.7%	89.6%	87.2%
Apartment & Office	89.6%	95.8%	89.3%	90.0%
Bar & Restaurant	59.5%	81.5%	57.8%	68.3%
Grocery & Supermarket	72.3%	72.5%	77.0%	76.2%
Clinic (Dental)	78.8%	80.7%	80.3%	78.1%
Outdoor	88.9%	90.6%	90.7%	87.7%

same IRB approval as the watch study, and the participants were allowed to keep the phones away or turn them off for a short period of time as needed for privacy reasons.

After the study was completed, the audio data was transferred to a server for annotation. The annotation scheme was the same as in the semi-naturalistic study, except that we only annotated the instances at a granularity of 30 seconds. For the one-week data collection, we found that the recording for one day of the week was shortened due to an unexpected battery issue. In total, we collected 5,421 30-second instances (45 hours) of audio, including 35 hours for the one-week data collection and 10 hours for the cross-subject data collection. Table 5 summarizes the main location contexts and associated activities of daily living captured in the entire study. The majority of contexts include office, apartment, vehicle, and street, which account for nearly 80% of the total. This is expected since the recording was mainly conducted during the daytime, where the majority of activities were performed in such places, including working, studying, having meetings, or walking outdoors.

6.2 Results of Free-living Study

To evaluate the performance of our conversation detector in the free-living study, we selected a checkpoint model (pruned) from the semi-naturalistic study and performed inference on the free-living data. Overall, the 3-class macro F1 score was 78.0%, with a precision of 74.5% and a recall of 83.4%. For the conversation class specifically, the class F1 score was 83.3%. Since we addressed the inference of user conversations, and face-to-face conversations were much more commonly captured than monologues in free-living situations, we reported conversation detection as a binary task in detail by grouping the location contexts (Table 6). We additionally reported the weighted average of the class F1 score, since this gives a sense of how the model performs for the majority of data collected from the unconstrained situations.

From Table 6, we can see that the model inference performance in the free-living study is promising for most scenarios. Also, the performance of the pruned model is still comparable to that of the original baseline without pruning. By examining the location contexts, we can see that the model shows a strong performance for instances collected from the apartment and office, and it performs reasonably well for outdoor situations as well. The model performs the worst for crowded scenarios, such as in a bar or a restaurant. This is expected because of the loud background noise and the overlap between the background voice activities and the user's conversations. To further investigate, we also computed the false positive and false negative rates in crowded scenarios. For user activities such as having a meal together or sitting in a cafe, the false positive and false negative rates were 17.6% and 16.4%, respectively. However, the false positive rate increased to 81% for conversation detection in a bar. A possible explanation is that the loud talking and singing voices in close proximity to the user significantly

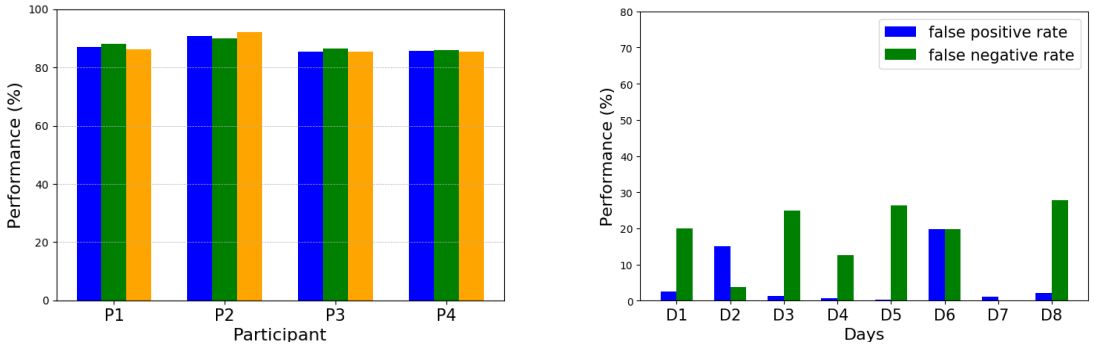

Fig. 5. Conversation detection performance for the free-living study across different tested subjects (left) and days (right). We examined the macro F1 (blue), recall (green), and precision (orange) on individual participants (left), and the false positive / negative rates per day to understand the model's robustness to varying human artifacts and environmental factors.

suppressed the user's own voice, which also confused the model in differentiating the foreground speech from the background sounds.

6.3 Result Analysis

To better understand the model detection performance at an event level, we examined the detection accuracy for the conversation class based on both 30-second instances and longer acoustic segments. In the free-living study, the class accuracy for the 30-second instances was 84.9%. To evaluate the performance for a conversational segment, we then grouped the ground truth instances if their adjacent instances belonged to the same conversations. A conversational segment was considered to be detected by the model if at least one of the instances inside the segment was correctly detected, and this shows the ability of our model to identify long conversational events. In this setup, the segment-level class accuracy was 92.0%, which shows a strong performance of our model to identify such long events. As a further step, we estimated the proportion of conversational instances that were correctly detected by our model within each conversational segment, and this can be used to estimate the duration of conversations. We obtained an accuracy of 81.3% for this test, which means that the majority of the duration of a conversation can be correctly identified by applying our model.

Figure 5 visualizes the detection performance of our model across human subjects and days of the study. From Figure 5 (left), we can see that the cross-subject performance of the model remains consistent, indicating the generalizability of our model against the variations of the users' behavioral patterns. From Figure 5 (right), we notice that the conversation detector tends to have a high false negative rate for the week-long test. By listening back to the collected audio, we found that conversational events of very short duration (e.g., a single back and forth) might be a possible reason for this. One possible solution is to include a dynamic windowing process to better adjust the model to these short events.

7 REAL-TIME SYSTEM IMPLEMENTATION

To demonstrate the effectiveness of our method in practical scenarios, we built a prototype of the system on a popular commercial smartwatch (Google Fossil Watch Gen 5) that has the same processor as the Fossil watch used in the semi-naturalistic study. The watch is equipped with a Snapdragon 3100 processor with 8 GB storage and 1 GB RAM. We replicated the feature extraction and inference steps in an Android application using Java to

Table 7. Comparison of the original and the optimized fusion model architectures regarding model sizes and inference latency per instance on a commodity Fossil smartwatch.

Model Setup	Model Size	Inference Latency
Original fusion model	3.0 MB	1,405 ms
Optimized fusion model	**0.8 MB**	**981 ms**

deploy the framework and evaluated the battery and memory consumption of the application on the smartwatch as discussed in this section.

7.1 Inference

Consistent with our previous studies, the deployed application is capable of inferring sound classes of *conversation*, *other speech*, and *ambient sound* at a granularity of 30 seconds. We define a cycle of inference as the entire process of recording audio, extracting features, and model inference. The application executes individual inference cycles independently when requested via the user interface. Once running, the application first triggers the device's microphone to continuously record audio at 16 kHz for 30 seconds. The FFT features are then calculated with the Noise[4] library, and the fusion network is called to generate the inference results. As discussed earlier, the deep learning models were validated with the Python TensorFlow library on a server; consequently, the model in the Android application is a selected checkpoint model from the semi-controlled study that is converted to TensorFlow Lite (TFLite). To reiterate, the 3-class inference result for the pruned model is 77% in the semi-controlled study, comparable to its baseline result (76%). Here we study the run-time capabilities of the TFLite models. Table 7 shows the comparison of checkpoint sizes in TFLite format with and without further optimization. For the pruned model, the TFLite default post-training optimization was also applied. As compared in the table, the size of the pruned TFLite model is only a quarter of the original model without further optimization, which is even more lightweight to be deployed on the watch. Besides, the average latency for FFT calculation and model inference is 981 ms per instance with the optimized checkpoint over ten continuous cycles of inference on the Fossil watch. This is also more preferable than the original fusion model without further optimization.

All processing was local on the device, and the recorded audio was deleted after each inference cycle. Currently, the execution of audio recording and FFT extraction / model inference is in series for research and evaluation purposes, and such latency is small compared to the input data duration of 30 seconds. In the future, we plan to run these modules in parallel so that extra latency can be eliminated.

7.2 Power Consumption

We profiled the battery usage of the application on the smartwatch to determine its baseline performance and identify areas in which the power consumption could be optimized. We monitored the battery level of the watch, which has a capacity of 310 mAh, while one user wore the smartwatch with the application running continuously until the device ran out of battery. Logs of the user's activities during these study periods were kept. During the study, the user performed a variety of activities in daily living, including attending class, working from home and in the office, socializing with other human subjects, exercising, cooking, and driving. As the baseline, the application performed individual inference cycles at a sampling rate of 16 kHz for each 30 second instance of audio in a continuous loop without any system- or power- level optimizations. A fully-charged watch ran for an average of 4.5 hours under this baseline design. Although this application could already be deployed to commodity smartwatches in real life to collect and analyze a reasonable amount of conversational data, we

[4] https://github.com/paramsen/noise

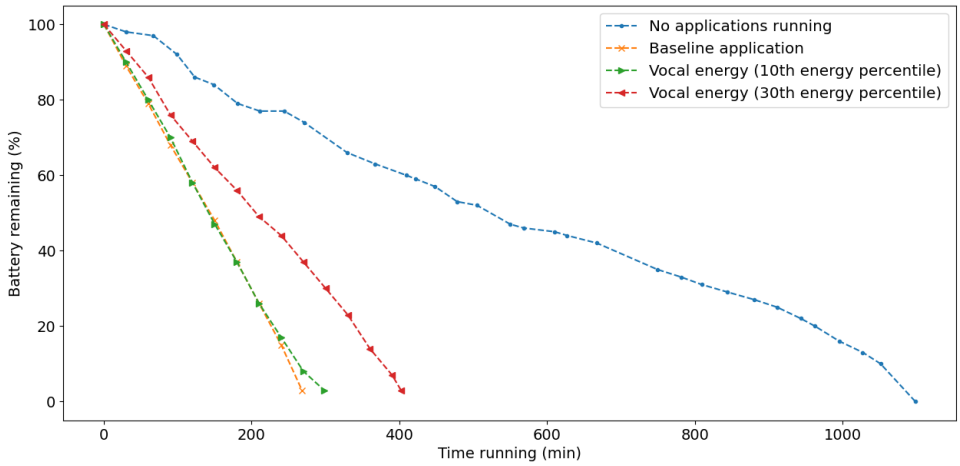

Fig. 6. Comparison of battery consumption by testing with different power setups. The tests were run on a Fossil watch with a 310 mAh battery.

still consider the optimization of its power consumption as an important step for better practicality. Therefore, we explored an audio energy-dependent adaptive sampling strategy to mitigate the power consumption of our application.

7.2.1 Audio Energy-dependent Adaptive Sampling. With the observation that the feature extraction and model inference processes are the most power-consuming processes of the application, we explored an energy-based VAD method to gate feature extraction and model inference without significantly missing conversations. The goal of this power optimization strategy is to detect periods of silence in which these processes do not need to be executed. To achieve this goal, the application first records one second of audio at 4 kHz, the lowest sampling rate supported by the smartwatch, and examines the accumulated energy of the signal. If the audio energy exceeds a threshold, the sampling rate increases to 16 kHz, and an inference cycle is triggered. Otherwise, the application proceeds with the lower sampling rate and continues in this conditional manner. We empirically compared two thresholds using data from the semi-naturalistic study based on the 10th and 30th percentile audio energy levels in one-second clips with conversations. With this strategy, the battery life lasted for 4.7 hours and 6.7 hours, respectively. Clearly, the strategy works better for a higher energy threshold, but this also comes at a cost of missing certain audio levels of conversations, which needs to be balanced in deployment.

7.2.2 Trade-offs between Conversation Detection and the Power Cost. Figure 6 compares the baseline application's battery usage with that of the adaptive sampling strategy. The plot also shows the battery life for a fully-charged watch of the same model without any applications running (18.3 hours). Admittedly, our current version of the application consumes a significant amount of power on the watch. However, as presented in Figure 6, the power consumption of our application can be mitigated with certain levels of optimization. We will discuss additional feasible power mitigation strategies in section 9.3. Given the importance of spoken communications in people's social lives and the close relationship between a user's conversational behaviors and mental wellness, the benefit of our system lies in the fact that it helps users to uncover and monitor these processes in a completely unconstrained and unobtrusive manner in daily living. With the current level of power optimization, our system can already be applied to several real-world scenarios to capture a meaningful duration of conversational behaviors with light

system burden. In the future, we hope to further deploy our application on dedicated devices so that its runtime can be extended to longitudinal use cases.

7.3 Runtime Overhead

In order to estimate the application overhead on the watch CPU and memory, we analyzed run-time information of the application extracted using Android *dumpsys* tool[5]. This tool provides statistics on system services running on connected devices, including the run time of every application, instantaneous device memory usage, memory usage over time, and CPU information per system service. We profiled the average proportional set size (PSS), a measure of RAM that a process uses, including overlapping memory shared with other processes, of the application over its entire run-time over one full battery charge. With the normal functions of the watch left untouched, e.g., with internet and Bluetooth connection, with system monitoring applications running, etc., the average PSS of the application was 41 MB, with a minimum of 37 MB and a maximum of 50 MB. The average PSS of the application accounted for an average of 4.2% memory usage of the entire watch system. The CPU load for the audio recording process was generally low (< 5%), but the feature calculation and model inference processes were much heavier, taking up nearly 50% of CPU usage.

8 HIGHER-ORDER SOCIAL INTERACTION APPLICATIONS

Our long-term goal is to develop a platform for quantifying face-to-face social interactions in real-world settings. As previously stated in the introduction, such a platform would support new applications in several disciplines, including behavioral sciences [14, 26] healthcare [63, 69] and social network analysis [11, 73]. To understand the feasibility of characterizing various aspects of social interactions using our automated conversation detection approach as a foundation, we built and evaluated three additional supervised models:

- **Social Context Recognition**: Once a conversation was detected, we tested a model to track the location and potential type of interaction, e.g., if the interaction happened during an office visit or at a party.
- **Substantive Conversation Detection**: As discussed by Basu [8], features of conversation scenes can be used to describe different types of face-to-face human conversations. In this setting, long conversation scenes typically indicate substantive conversations, which have been shown to be associated with greater well-being [55]. We tested a model to characterize conversation type by its duration.
- **Social Engagement Recognition**: During a social interaction, there is a strong correlation between turn-takings and a speaker's engagement in the conversation [12, 24]. We tested a model to explore whether we could quantify conversation engagement from detected conversations.

8.1 Modeling

To develop the supervised models, we leveraged five days of the free-living dataset, containing 1,698 minutes (3,395 instances) of 30-second audio snippets. Among the instances, 205 minutes were detected as conversations by our system. The output instances were annotated by a researcher based on the above three categories (social context, conversation scene, social engagement). Specifically, we targeted four contexts based on the encountered conversational events, including *outdoor*, *bar*, *office and grocery*, and *others*. Class *outdoor* was for events outside of a building or in a vehicle. Class *bar* was associated with social events in a bar or restaurant. *Office and grocery* included events of dental visits, conversations in a leasing office, or in a grocery. Other conversational events or instances that failed to be detected by the conversation detection system were categorized as *others*. The conversation scene was divided into two groups. The first group referred to conversations of long duration (no less than one third of the instance length of 30 seconds), and the others were annotated as short scenes. Similarly, two groups were created for the engagement study. Detected conversations with only a single back and forth

[5] https://developer.android.com/studio/command-line/dumpsys#syntax

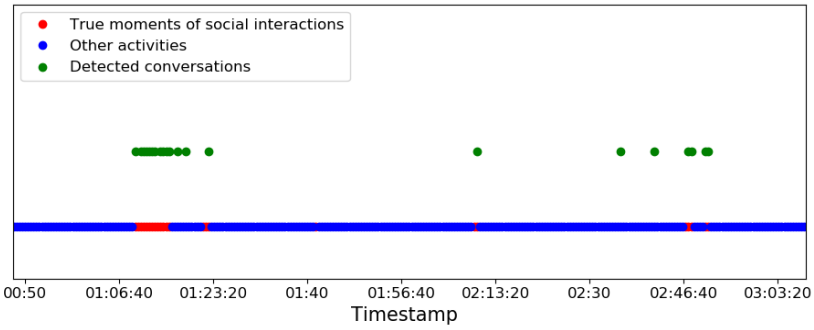

Fig. 7. Diary of face-to-face conversational moments generated based on the conversation time stamps logged by our app. This information supports a full understanding of when and how long the device wearer physically interacts with other human subjects within a given time window.

by the speakers or no switch of speaker turns were annotated as low engagement. Otherwise, the conversation was annotated as high engagement. Examples of low engagement in our dataset included short greetings and instructions with short responses such as 'yes' or 'okay'. In total, we collected 114 minutes of conversations with long scenes and 91 minutes otherwise. 116 minutes of conversations were categorized as high engagement, and the remaining 89 minutes were categorized as low or no engagement, including non-conversational instances incorrectly detected by the conversation detection system.

The supervised models we used for the study were based on hybrid convolution and fully-connected neural networks. We built separate models for each of the recognition tasks. The architecture of the models is as follows:

Input → Conv2D[32] → Conv2D[64] → Conv2D[64] → Flatten → FC[128] → FC[64] → FC[X]

where Conv2D[K] denotes a convolutional layer with K channels, each followed by a 2D max pooling of (2×2) pool size and strides. The kernel size of each convolutional layer was (3×3), with a stride of 1 for the first two layers and 2 for the third. FC[K] denotes a fully-connected layer of size K activated by the ReLU activation. The output dimension X was chosen based on the number of target classes. The model was fed with the same 2D spectrogram features used for conversation detection. Besides, the parameter size of the models was 0.5M, which is lightweight to be deployed on smartwatches in addition to the current application.

8.2 Results and Analysis

The models were developed and evaluated based on a 2-fold cross validation scheme, and the average of the best fold-level performance was reported. By examining the models trained for context recognition, we observed that most conversational contexts captured outdoors or in the bar / restaurant were correctly recognized. This is expected, since the background noise in such contexts, including wind sounds and background crowd noise, are often unique. Overall, the models achieved 85.1% macro F1 score, with a macro precision / recall of 85.4% / 85.1%. The models developed for scene and engagement detection also showed reliable performance. To detect conversations of long scenes, the models achieved 82.2% macro F1 score, with a macro precision and recall score of 82.3% and 82.2%, respectively. The F1 score for detecting high social engagement was 83.4%, with the same precision and recall values. By deriving these cascaded models, we can further characterize detailed information of the wearer's conversations based on the detected conversational instances. For example, we may infer that the wearer is having a formal meeting or discussion with others if several conversations of long scenes and high engagement are detected in an office.

In addition to the above models, the time stamps of the detected conversational instances can also be logged by our system. By gathering this information, we further show an example diary of the wearer's detected conversations for a 3-hour time window of the free-living study (Figure 7). In the figure, it can be seen that there is a strong temporal correlation between the logged conversational events and face-to-face social interactions. Hence, this information supports a detailed diary of the wearer's social moments, which can be used to automatically track when and how long the wearer physically interacts with other human subjects within the detection window.

9 DISCUSSIONS

In this section, we discuss additional comparisons and analyses that are relevant when considering real-world deployment of our proposed system: privacy, power consumption, the impact of clothing on data collection, and how data collection and subsequent performance differ between a smartphone and a smartwatch.

9.1 Smartphone vs. Smartwatch

As described in Section 6.1, we collected audio data with a smartwatch and a smartphone for three participants in the free-living study. The objective of introducing smartphones in the study was to understand how automated face-to-face conversation detection might differ between these two devices in real life conditions. Smartphones are highly personal devices but they are not *always* as close to individuals as a wrist-strapped smartwatch, impacting the quality and relevance of captured audio. For external validity, participants were instructed to carry the smartphone similarly to how they did with their own personal phones. In data analysis, we followed the same steps to extract the spectrogram and foreground speech features from the smartphone data as we did with the smartwatch. Finally, we performed model inference with the smartphone data using the same checkpoint model as in the free-living study and compared the results against those with the smartwatch data.

As expected, participants did not keep the phone immediately close to them at all times, which affected the quality of the audio. Our results showed that the macro F1 score for binary conversation detection dropped to 78.1% with the smartphone data for the participants (89.1% for the watch test). Similarly, the macro precision and recall dropped to 72.7% and 84.3%, respectively (88.7% and 89.5% for the watch test). To further investigate this aspect, we examined the precision for individual participants, and the precision values of the watch / phone tests were 96.3% / 71.8%, 76.6% / 60.4%, and 93.6% / 87.4%, respectively. Based on the participants' logs, the most probable factor that affected the model's performance for smartphones was that the participants mostly placed their phones in their pockets or handbags during the test rather than always leaving them in hand or on the wrist as with their watches. When placed in a pocket or handbag, the audio recorded by the smartphone was inevitably affected by the surrounding material and became muffled or dampened. In some instances, this resulted in missing conversations entirely, which caused a drop in recall.

9.2 Smartwatch Data Collection Underneath Clothing

Our free-living study was conducted in the spring and summer months, and participants were dressed for warm weather. Therefore, the smartwatch was never hidden away underneath a sleeve or extra layers of clothing. However, this is a realistic concern in winter months or colder climates. To explore this scenario, we overlaid the audio captured in the free-living study with acoustic artifacts that would be present in this setting, such as the sound of clothing rubbing against the smartwatch. To do so, one user wore a device on a wrist with a sleeve covering and collected such acoustic artifacts for 30 continuous seconds. We then overlaid the dataset and the acoustic artifacts at multiple signal-to-noise levels (3 dB and 20 dB) to simulate thin and thick layers of fabric covering. The macro F1 score for the 3-class inference was 73.0% and 77.5%, respectively, for the two covering levels (78.0% without covering, as presented earlier). For the detection of conversations, specifically, the binary macro F1 score was 87.5% and 89.4%, respectively, where it was 89.2% without covering. Based on these results,

we confirmed that while the model's performance slightly decreased in the covered setting, it was also robust to such acoustic artifacts.

9.3 Context-Driven Power Optimizations

Our application can run for at least 4.5 hours on a commodity smartwatch while simultaneously and continuously collecting audio data and performing conversation detection. Given the limited battery capacity of smartwatches, this is a promising achievement. However, it would be desirable to extend the operational range of the application to at least 8-10 hours in order to approximate an entire workday. In section 7.2.1, we described a technique for reducing power usage by dynamically adjusting the sampling rate. Here, we revisit the topic of power savings by presenting early findings on two techniques that block the smartwatch's microphone from capturing audio based on user-specific contextual information. Beyond power reduction, a desirable byproduct of these techniques is that they mitigate privacy concerns by constraining audio capture and processing for designated locations and / or activities.

9.3.1 Location Context. A user's location can be an important determining factor in deciding whether the system should be fully operational or not. In locations where conversations are not likely to take place, e.g., in a movie theater or at a religious service, it might be reasonable to dynamically turn off audio capture to extend the device's battery life. Such assumptions will sometimes lead to inaccurate results, e.g., if conversations take place before or after a movie while an individual is still at the movie theater. In this case, there is a clear trade-off between battery life and conversation sensitivity. An alternative approach might be a user-configurable setting for restricting conversation detection to certain places, e.g., work or home. For example, if a person lives alone and does not typically interact with others after work, then the application can be programmed to turn off the microphone and block audio capture when the user's home WiFi network is detected. To experiment with this technique, we conducted an initial proof-of-concept study where an individual who lives alone configured the application to turn off the microphone when his *home* WiFi network was detected. Under these conditions, the battery of the watch was extended to 6.25 hours when the individual was at home for 1.5 hours, and 9 hours when the individual was at home for 6.5 hours. Although the gains from this technique depend on the amount of time at predefined locations, our preliminary results validate this approach as a way to reduce the application's power consumption and mitigate audio privacy concerns.

9.3.2 Activity Context and Wearability. Another contextual element that can be used to determine the likelihood of conversations, and whether audio should be captured and processed continuously, is the type of activity an individual is performing. For example, while swimming, reading or meditating, it is very unlikely that conversations will take place. Another useful cue for determining whether to trigger the audio processing pipeline is whether the individual is actually wearing the smartwatch. This can be reliably achieved with sensors present in smartwatches today. To demonstrate the feasibility of this technique, we conducted a pilot study in which one individual launched the application in the evening and went to bed without wearing the watch. The smartwatch, with the application running, was operational for 13.9 hours, with the watch off-body during 10.5 hours of that duration. Again, while the gains of this technique depend on the duration of the watch off-body, this initial result shows how physical activities where conversations are unlikely can be used to extend the application's battery life by reducing the duration of audio processing. Undeniably, if the user takes the smartwatch off for additional activities that do involve conversations, conversations will be missed, so this microphone-gating technique should only be applied to activities for which there is high confidence that few conversations will take place. This is also a trade-off between the application's runtime and conversation sensitivity that is intrinsic to this power mitigation technique.

9.4 Privacy Considerations

Privacy concerns are front and center when it comes to audio capture in naturalistic and conversation settings [32, 44, 78]. The primary way our current implementation addresses this issue is by executing the entire capture and inference processes *on board*, without transmitting or saving raw audio. Additionally, we also explored techniques that result in privacy mitigation by triggering audio capture dynamically, depending on contextual factors such as location, activity and device wearability. Yet, despite these measures, additional steps for protecting privacy can be taken. In the future, we plan to extend our system such that the smartwatch can not only capture sensor data but also filter or degrade the intelligible information within the audio on-board. As such, inference will run on less sensitive data from users [62], thereby further limiting the risks of privacy leakage.

10 LIMITATIONS AND FUTURE DIRECTIONS

While the research presented in this paper points in a promising direction, it is important to underscore its limitations and discuss opportunities for the future. Starting with our proposed approach and real-time system, we explored how different power optimization strategies may be applied to improve the practicality of our application. In addition to those approaches, there is still room to improve the model design and application implementation. For example, our current solution addresses the potential connection between the on- / off-body status of a smartwatch and the user's expected conversations. In many real-life scenarios, there can be a broader set of non-verbal behavioral cues, such as motion patterns of the wrist, which may be used as indicators to reveal a user's face-to-face conversations. Secondly, with our WiFi detection, one may set up a customized list of locations based on a user's conversation routines, and a further step is to build an active learning pipeline to dynamically adjust the gating contexts based on the user's preference. Thirdly, the usage of acoustic spectrogram features may be further optimized in our system. Currently, our FSD module is executed in parallel with the remaining part of the detection network. A possible direction is to enable a cascaded design where the FSD can be executed first to examine the possible existence of user speech before triggering the extraction of the spectrogram features. This is meaningful because the foreground detection process may be enabled based on low-level features alone [56], and the extraction of such features can be less power-consuming. We intend to address such directions as future work to further improve our application.

Additionally, our framework can still be improved according to different scenarios. For example, we focused on instance-level recognition in our study with a fixed instance size of 30 seconds. However, human speech and conversations can be highly dynamic in real life, and a fixed window size may fall short when it comes to further characterizing conversations, such as revealing the moments of turn-taking behaviors [6]. Also, the inference performance for *other speech* may be improved by leveraging a smaller window, especially given monologues often short in duration. As mentioned earlier, our choice was mostly inspired by prior work in conversational analysis. A possible future way to improve the model could be the development of adaptive windows based on actual user speech.

11 CONCLUSION

Automated detection of physical conversations in daily living is helpful for quantifying user behaviors regarding social interactions and potentially revealing the user's mental health states. This paper presented a practical system for user conversation detection with a single smartwatch by leveraging acoustic sensing. Based on our neural network-based framework, our system was able to infer the existence of user physical conversations and user speech against ambient sounds. To evaluate the framework, we conducted a semi-naturalistic study based on 18 groups of participants in their own homes and a free-living study based on four participants wearing a commercial smartwatch for continuous audio recording, including one participant collecting data over a week. For detecting conversations, our model achieved 83.2% F1 score in the semi-naturalistic study and 83.3% F1 score

in the free-living study. We further discussed in depth the detailed performance of the proposed framework, comparisons of different modeling methods, and a prototype of the system to demonstrate its run-time capability on a popular smartwatch.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Rebecca Adaimi, Howard Yong, and Edison Thomaz. 2021. Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–24.

[3] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

[4] Mohsin Y Ahmed, Sean Kenkeremath, and John Stankovic. 2015. Socialsense: A collaborative mobile platform for speaker and mood identification. In *European Conference on Wireless Sensor Networks*. Springer, 68–83.

[5] Apple. 2023. AirPods redefine the personal audio experience. https://www.apple.com/newsroom/2023/06/airpods-redefine-the-personal-audio-experience/. Accessed: 2023-07-13.

[6] 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. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 1 (2018), 1–27.

[7] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[8] Sumit Basu. 2002. *Conversational scene analysis.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[9] Gary S Becker. 1974. A theory of social interactions. *Journal of political economy* 82, 6 (1974), 1063–1093.

[10] 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*. 10–19.

[11] Yan-Ann Chen, Ji Chen, and Yu-Chee Tseng. 2015. Inference of conversation partners by cooperative acoustic sensing in smartphone networks. *IEEE Transactions on Mobile Computing* 15, 6 (2015), 1387–1400.

[12] Tanzeem Khalid Choudhury. 2004. *Sensing and modeling human networks.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[13] Keum San Chun, Sarnab Bhattacharya, and Edison Thomaz. 2018. Detecting eating episodes by tracking jawbone movements with a non-contact wearable sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–21.

[14] Agnete Skovlund Dissing, Tobias Bornakke Jørgensen, Thomas Alexander Gerds, Naja Hulvej Rod, and Rikke Lund. 2019. High perceived stress and social interaction behaviour among young adults. A study based on objective measures of face-to-face and smartphone interactions. *PloS one* 14, 7 (2019), e0218429.

[15] Susan Kay Donaldson. 1979. One kind of speech act: How do we know when we're conversing? (1979).

[16] Andreas Ejupi and Carlo Menon. 2018. Detection of talking in respiratory signals: A feasibility study using machine learning and wearable textile-based sensors. *Sensors* 18, 8 (2018), 2474.

[17] Tiantian Feng, Amrutha Nadarajan, Colin Vaz, Brandon Booth, and Shrikanth Narayanan. 2018. Tiles audio recorder: an unobtrusive wearable solution to track audio activity. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*. 33–38.

[18] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.

[19] Google. 2020. Sound classification with YAMNet. https://www.tensorflow.org/hub/tutorials/yamnet. Accessed: 2022-08-01.

[20] Hall and Watson. 1970. NASA Exercise: Survival on the Moon. https://www.humber.ca/centreforteachingandlearning/assets/files/pdfs/MoonExercise.pdf. Accessed: 2020-10-01.

[21] Judith Holler, Kobin H Kendrick, Marisa Casillas, and Stephen C Levinson. 2016. *Turn-taking in human communicative interaction.* Frontiers Media SA.

[22] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 493–504.

[23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).
[24] Joey Chiao-yin Hsiao, Wan-rong Jih, and Jane Yung-jen Hsu. 2012. Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.
[25] William Huang, Ye-Sheng Kuo, Pat Pannuto, and Prabal Dutta. 2014. Opo: a wearable sensor for capturing high-fidelity face-to-face interactions. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*. 61–75.
[26] Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2020. Can Prediction of Turn-management Willingness Improve Turn-changing Modeling?. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.
[27] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2704–2713.
[28] Kleomenis Katevas, Katrin Hänsel, Richard Clegg, Ilias Leontiadis, Hamed Haddadi, and Laurissa Tokarchuk. 2019. Finding dory in the crowd: Detecting social interactions using multi-modal mobile sensing. In *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems*. 37–42.
[29] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. 2013. Multisensor data fusion: A review of the state-of-the-art. *Information fusion* 14, 1 (2013), 28–44.
[30] Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. 2008. Meeting mediator: enhancing group collaborationusing sociometric feedback. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 457–466.
[31] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[32] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring privacy concerns about personal sensing. In *International Conference on Pervasive Computing*. Springer, 176–183.
[33] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.
[34] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
[35] 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.
[36] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 213–224.
[37] Gierad Laput and Chris Harrison. 2019. Sensing fine-grained hand activity with smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[38] Kornel Laskowski, Mari Ostendorf, and Tanja Schultz. 2007. Modeling vocal interaction for text-independent classification of conversation type. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*. 194–201.
[39] 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. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 375–388.
[40] Haochao Li, Eddie CL Chan, Xiaonan Guo, Jiang Xiao, Kaishun Wu, and Lionel M Ni. 2015. Wi-counter: smartphone-based people counter using crowdsourced wi-fi signal data. *IEEE Transactions on Human-Machine Systems* 45, 4 (2015), 442–452.
[41] Qiang Li, Shanshan Chen, and John A Stankovic. 2013. Multi-modal in-person interaction monitoring using smartphone and on-body sensors. In *2013 IEEE International Conference on Body Sensor Networks*. IEEE, 1–6.
[42] Dawei Liang, Guihong Li, Rebecca Adaimi, Radu Marculescu, and Edison Thomaz. 2022. Audioimu: Enhancing inertial sensing-based activity recognition with acoustic models. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*. 44–48.
[43] Dawei Liang, Yangyang Shi, Yun Wang, Nayan Singhal, Alex Xiao, Jonathan Shaw, Edison Thomaz, Ozlem Kalinli, and Mike Seltzer. 2021. Transferring Voice Knowledge for Acoustic Event Detection: An Empirical Study. *arXiv preprint arXiv:2110.03174* (2021).
[44] 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*. 1–10.
[45] Dawei Liang and Edison Thomaz. 2019. Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–18.
[46] Bethany Little, Ossama Alshabrawy, Daniel Stow, I Nicol Ferrier, Roisin McNaney, Daniel G Jackson, Karim Ladha, Cassim Ladha, Thomas Ploetz, Jaume Bacardit, et al. 2020. Deep learning-based automated speech detection as a marker of social functioning in late-life depression. *Psychological Medicine* (2020), 1–10.
[47] Shangqing Liu, Yanchao Zhao, and Bing Chen. 2017. WiCount: A deep learning approach for crowd counting using WiFi signals. In *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*. IEEE, 967–974.

[48] Hong Lu, AJ Bernheim Brush, Bodhi Priyantha, Amy K Karlson, and Jie Liu. 2011. Speakersense: Energy efficient unobtrusive speaker identification on mobile phones. In *International conference on pervasive computing*. Springer, 188–205.
[49] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 351–360.
[50] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2009. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. 165–178.
[51] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2010. The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM conference on embedded networked sensor systems*. 71–84.
[52] Chengwen Luo and Mun Choon Chan. 2013. Socialweaver: Collaborative inference of human conversation networks using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. 1–14.
[53] Aleksandar Matic, Venet Osmani, Alban Maxhuni, and Oscar Mayora. 2012. Multi-modal mobile sensing of social interactions. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. IEEE, 105–114.
[54] Matthias R Mehl, James W Pennebaker, D Michael Crow, James Dabbs, and John H Price. 2001. The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior research methods, instruments, & computers* 33, 4 (2001), 517–523.
[55] Matthias R Mehl, Simine Vazire, Shannon E Holleran, and C Shelby Clark. 2010. Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological science* 21, 4 (2010), 539–541.
[56] Amrutha Nadarajan, Krishna Somandepalli, and Shrikanth S Narayanan. 2019. Speaker agnostic foreground speech detection from audio recordings in workplace settings from wearable recorders. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6765–6769.
[57] Annamalai Natarajan, Deepak Ganesan, and Benjamin M Marlin. 2019. Hierarchical active learning for model personalization in the presence of label scarcity. In *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 1–4.
[58] Niklas Palaghias, Seyed Amir Hoseinitabatabaei, Michele Nati, Alexander Gluhak, and Klaus Moessner. 2015. Accurate detection of real-world social interactions with smartphones. In *2015 IEEE International Conference on Communications (ICC)*. IEEE, 579–585.
[59] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 385–394.
[60] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 281–290.
[61] Md Mahbubur Rahman, Amin Ahsan Ali, Kurt Plarre, Mustafa Al'Absi, Emre Ertin, and Santosh Kumar. 2011. mConverse: Inferring conversation episodes from respiratory measurements collected in the field. In *Proceedings of the 2nd Conference on Wireless Health*. 1–10.
[62] Andrew Raij, Animikh Ghosh, Santosh Kumar, and Mani Srivastava. 2011. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 11–20.
[63] Marianne Schmid Mast, Daniel Gatica-Perez, Denise Frauendorfer, Laurent Nguyen, and Tanzeem Choudhury. 2015. Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science* 24, 2 (2015), 154–160.
[64] Abhishek Sehgal and Nasser Kehtarnavaz. 2018. A convolutional neural network smartphone app for real-time voice activity detection. *IEEE Access* 6 (2018), 9017–9026.
[65] Deepa Sethi and Manisha Seth. 2009. Interpersonal communication: Lifeblood of an organization. *IUP Journal of Soft Skills* 3 (2009).
[66] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. 2011. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one* 6, 8 (2011), e23176.
[67] Shaojie Tang, Jing Yuan, Xufei Mao, Xiang-Yang Li, Wei Chen, and Guojun Dai. 2011. Relationship classification in large scale online social networks and its impact on information propagation. In *2011 Proceedings IEEE INFOCOM*. IEEE, 2291–2299.
[68] Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 1029–1040.
[69] Weichen Wang, Shayan Mirjafari, Gabriella Harari, Dror Ben-Zeev, Rachel Brian, Tanzeem Choudhury, Marta Hauser, John Kane, Kizito Masaba, Subigya Nepal, et al. 2020. Social Sensing: Assessing Social Functioning of Patients Living with Schizophrenia using Mobile Phone Sensing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[70] Martin Warren. 2006. *Features of naturalness in conversation*. Vol. 152. John Benjamins Publishing.
[71] John Wilson. 1987. On the topic of conversation as a speech event. *Research on Language & Social Interaction* 21, 1-4 (1987), 93–114.
[72] Danny Wyatt, Tanzeem Choudhury, and Jeff A Bilmes. 2007. Conversation detection and speaker segmentation in privacy-sensitive situated speech data.. In *Interspeech*. 586–589.

[73] Danny Wyatt, Tanzeem Choudhury, and Henry Kautz. 2007. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4. IEEE, IV–213.

[74] 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 *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 43–52.

[75] Zhixian Yan, Jun Yang, and Emmanuel Munguia Tapia. 2013. Smartphone bluetooth based social sensing. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 95–98.

[76] Yinxue Yi, Zufan Zhang, Laurence T Yang, Xianjun Deng, Lingzhi Yi, and Xiaokang Wang. 2020. Social interaction and information diffusion in Social Internet of Things: dynamics, cloud-edge, traceability. *IEEE Internet of things Journal* 8, 4 (2020), 2177–2192.

[77] Ruiqing Yin, Hervé Bredin, and Claude Barras. 2017. Speaker change detection in broadcast tv using bidirectional long short-term memory networks. In *Interspeech 2017*. ISCA.

[78] Pablo Pérez Zarazaga, Sneha Das, Tom Bäckström, Vishnu Vidyadhara Raju Vegesna, and Anil Kumar Vuppala. 2019. Sound Privacy: A Conversational Speech Corpus for Quantifying the Experience of Privacy.. In *Interspeech*. 3720–3724.

[79] Huanle Zhang, Wan Du, Pengfei Zhou, Mo Li, and Prasant Mohapatra. 2016. DopEnc: acoustic-based encounter profiling using smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 294–307.

[80] Michael Zhu and Suyog Gupta. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878* (2017).