Harnessing WiFi Sensing for Video-less Monitoring of Participants' Engagement during Online Synchronous Meetings

VIJAY KUMAR SINGH*, IIIT Delhi, India
ADITYA MISHRA, IIIT Delhi, India
HIMANSHU SHEKHAR, IIIT Delhi, India
PRAGMA KAR, Kalinga Institute of Industrial Technology Deemed to be University, India
SANDIP CHAKRABORTY, IIT Kharagpur, India
MUKULIKA MAITY, IIIT Delhi, India

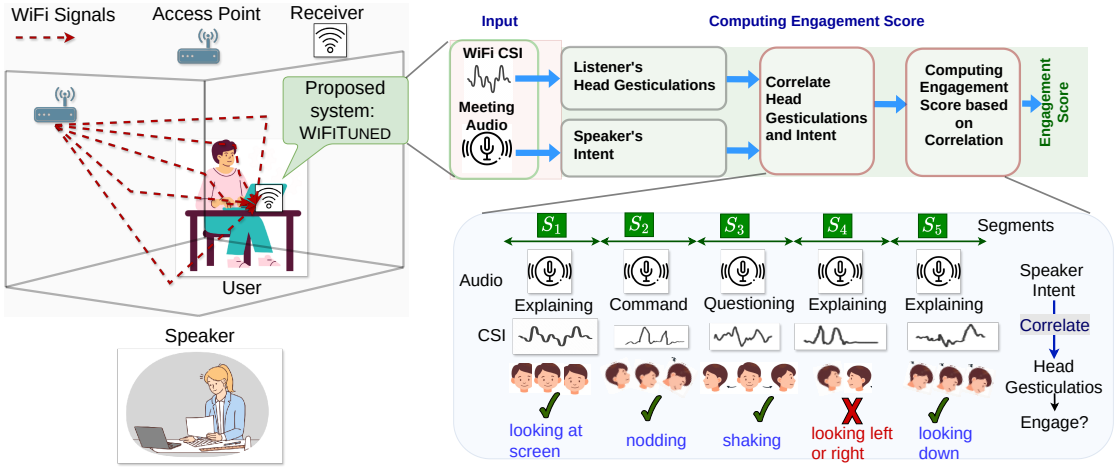


Fig. 1. Core idea of WiFiTuned: A user joins online meetings with WiFiTuned using a WiFi network in an indoor environment. WiFiTuned collects WiFi-CSI data and meeting audio and uses them as inputs. WiFiTuned divides the entire meeting into smaller 10*s* segments. For each segment, it detects the participant's head gesticulations and the speaker's intent through WiFi CSI and Audio, respectively. Then, it correlates the two to classify each segment as *engage* or disengage. Finally, the engagement score for the entire meeting is computed.

This paper proposes a multi-modal, non-intrusive, and privacy-preserving system WiFiTuned for monitoring participants' engagement for synchronous online meetings. It uses two sensing modalities, i.e., WiFi Channel State Information (CSI) and audio, captured over the

*

Authors' addresses: Vijay Kumar Singh, IIIT Delhi, New Delhi, New Delhi, India, vijaysi@iiitd.ac.in; Aditya Mishra, IIIT Delhi, New Delhi, India, aditya21125@iiitd.ac.in; Himanshu Shekhar, IIIT Delhi, New Delhi, India, himanshu21152@iiitd.ac.in; Pragma kar, Kalinga Institute of Industrial Technology Deemed to be University, Orissa, India; Sandip Chakraborty, IIT Kharagpur, West Bengal, India, sandipchkraborty@gmail.com; Mukulika Maity, IIIT Delhi, New Delhi, India, mukulika@iiitd.ac.in.

Manuscript submitted to ACM

meeting device (such as a laptop) to monitor the participants passively without using any visual information. At its core, WiFiTuned detects the head gestures of the participants through WiFi CSI and detects the speaker's intent through audio. Then, it correlates the two modalities to infer participants' engagement. We evaluate WiFiTuned with 22 participants. It detects the engagement level with an average accuracy of > 86%. The developed head gesticulation recognition model adapts to different participants' positions, locations, and environments by obtaining a classification accuracy of 91.68%. From a thorough usability study of the developed platform, we observed that WiFiTuned attains an average usability score of 79.73% on the system usability scale (SUS), indicating overall adaptability for real-world usage.

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

Additional Key Words and Phrases: Online meetings, Passive monitoring, WiFi sensing, Multimodal systems

ACM Reference Format:
Vijay Kumar Singh, Aditya Mishra, Himanshu Shekhar, Pragma kar, Sandip Chakraborty, and Mukulika Maity. 2018. Harnessing WiFi Sensing for Video-less Monitoring of Participants' Engagement during Online Synchronous Meetings. *J. ACM* 37, 4, Article 111 (August 2018), 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Virtual meetings and classrooms have become increasingly popular in recent years due to socioeconomic changes after the COVID-19 pandemic. Recent statistics[1] indicates that about 59% of the college students have taken some or all of their classes in 2021; although this rate reduced slightly further in subsequent years, but remained over 50%. As expressed in many surveys and the literature [20, 31, 55], one of the significant issues with synchronous online classrooms is the interaction gap among the students and the course instructors, mainly due to the absence of proper visual contacts. In addition, multitasking is common during synchronous online classes [32], where the students can take notes, search relevant content online, chat with friends, use mobile phones, or even take food/drink. While some types of multitasking, like taking notes or looking for relevant materials, can be positive and promote attentiveness, others (like chatting or browsing social media over mobile) can have negative impacts. In the absence of a proper monitoring system, such negative multitasking not only affects the student but also impacts the overall class as well as the pace of teaching by creating a communication barrier between the students and the instructors [5, 33]. Consequently, it is important to continuously monitor student engagement during synchronous online classes, which can help provide effective, timely instructor feedback.

However, the idea of monitoring user engagement during online meetings is not new; several prior works [6, 10, 18, 29, 33, 36] have demonstrated efficient approaches to monitor user engagement and even detected instances of multitasking during synchronous online meetings. These works primarily use the participant's video feed as the primary modality and utilize techniques like modelling eye gestures, body movements, head gestures, etc., or facial expressions to infer participants' attentiveness. However, video as a sensing modality might not be compelling as participants may prefer to keep their video off due to the apparent reason for privacy or poor bandwidth connectivity. Further, video-based inference significantly depends on the lighting condition, facial occlusion, camera angle, camera resolution, etc. Therefore, a ubiquitous platform for attentiveness monitoring needs a non-intrusive, passive, on-device scalable, usable, and privacy-preserving sensing modality to develop the solution.

This paper considers the fact that nowadays, WiFi is omnipresent in the indoor space, and most online participants utilize wireless connectivity at the last mile. Notably, in recent years, WiFi sensing using *Channel State Information*

[1]https://www.forbes.com/advisor/education/online-colleges/online-learning-stats/ (Last accessed: August 8, 2024)

(CSI) has emerged as one of the most promising areas for Human Activity Recognition (HAR) [35, 60] and Human Gesture Recognition (HGR) [17]. Exploiting the CSI captured over the end devices being used (such as mobile phones and laptops) and correlating it with the speaker's audio in the meeting, we develop a multi-modal system named WiFiTuned to pervasively monitor the participant's attentiveness in a non-intrusive, passive, scalable, and usable privacy-assured manner. Fig. 1 shows the core idea of WiFiTuned. Prior work [4] has shown that head gesticulations of an individual have evolved similarity with the intent of speech of the speaker. For example, a listener would essentially nod/shake their head in agreement/disagreement if they are paying attention to online meetings where the speaker makes a statement. Similarly, disengaged users, when browsing social media on mobile, are likely to show no response when the speaker gives a command. Through an anonymous survey of over 300 participants (having experience with online synchronous meetings) from four different countries, we hypothesize that head gesticulations correlate well with the engagement level during synchronous online meetings (details in §3). Moreover, head gesticulations align with the speaker's intent related to engagement and disengagement. WiFiTuned works on this hypothesis and correlates the listener's head gesticulation pattern with the speaker's intent to quantify the engagement level. The model uses two modalities - *CSI and audio*, CSI to recognize head gesticulations and audio to recognize the speaker's intent. Head gesticulations and intent are then utilized to infer the engagement levels of the participants.

Challenges: However, developing an automated engagement monitoring system has several challenges, as follows.

(1) Considering the notion of "*engagement*" is subjective, deciding on an objective evaluation for engagement is not straightforward. Existing literature [12, 43, 45] classifies engagement from three different contexts – (i) *Affective/emotional engagement*, (ii) *behavioral engagement*, and (iii) *cognitive engagement*. In this paper, we primarily focus on *behavioural engagement*, which indicates "*the extent to which individuals can be observed to exert effort and show persistence to remain involved in an activity or situation*" [12]. Following this definition of engagement, it is necessary to characterize the *behaviours* that signify their engagement during online synchronous meetings. Notably, such behaviour can vary across individuals depending upon their race, age, gender, local micro-cultures, and various other factors. Therefore, such behavioural cues must be chosen carefully to build a model.
(2) Participants attend online meetings from many different environments. The challenge is to develop an engagement monitoring system that will be robust across users and various environments. Notably, the WiFi CSI gets impacted by the environment and the subjects/objects therein. For example, CSI gets impacted by blockages like walls, the presence of other WiFi devices in the vicinity, the mobility of the user, etc. Therefore, it is important to develop a model which is robust against such factors.
(3) Developing a correlation model between the head gesticulations and the audio intent is challenging as we must provide quantitative ways of correlating them, and there might be multiple different head gesticulations corresponding to the same audio intent. For example, in an online class, when an instructor asks *whether the students have understood the concept explained*, some may nod while some may shake their head; however, both refer that they are engaged.

Contributions: Considering the above challenges, we develop a novel framework in this paper by modelling the head gesticulations inferred from the CSI data under diverse environments and correlating it with the speaker's intent through hierarchical clustering. Considering the observations from a thorough anonymous survey, we developed a head-gesticulation recognition model carefully, knowing that WiFi CSI is impacted by environmental changes. The extracted information from WiFi CSI has further been correlated with the audio intent to obtain an *engagement score* of

a participant for the entire meeting duration. In contrast to the existing works, the contributions to this paper are as follows.

(1) **Extracting environment-independent head gesticulation features from the CSI data:** Acknowledging the fact that WiFi CSI gets impacted by the obstacles and environments between the transmitter and the receiver, we develop a novel solution to monitor various head gesticulations in an environment-independent manner accurately. Our approach considers denoising the CSI data, followed by the extraction of a novel feature called the *Doppler Phase Vector*, which captures the impact of head gesticulations on the WiFi CSI captured over the device (laptop or mobile) on which the subject is working. Note that state-of-the-art WiFi sensing work [8, 35] collected WiFi CSI data in a controlled environment and thus performed only a single pre-processing step, i.e., *phase sanitation*. However, mere phase sanitization does not help extract true environment-independent features. Hence, we perform further denoising procedures such as applying various signal processing steps *Hampel filter, 1-D wavelet transform filter, and Savgol filter* to make it robust across various environments (details in §.4.2). We then use a novel deep learning-based approach to recognize the head gesticulation of the participants based on attention modeling to handle the environmental diversities. The proposed model achieves an average accuracy of 91.68% across different participants and locations in semi-controlled settings where the participants do not control the placement of the devices. However, to ensure robustness and practical usability, we also evaluate the model under in-the-wild settings that reflect real-world scenarios with various positioning of the participant and the WiFi access point (varied angles, distances, position, presence of obstacle, presence of interferer), crowded vs. empty rooms, and different sitting positions such as lying, leaning forward, and leaning backward. From thorough evaluation under different setups, we observe that the model demonstrates adaptability and achieves an average accuracy of 86%.

(2) **Extracting audio intent and correlate with head gesticulations to infer the engagement score:** To correlate head gesticulation and the speaker's intent from the audio, we divide the entire meeting into small-sized segments and group them in two hierarchies – first, based on the head gesticulation patterns and then on the speaker's intent. Next, we determine whether a sub-group and associated segments are *engaged* or disengaged by matching the actual head gesticulations with expected ones for that sub-group. Finally, based on individual segments' status, we obtain an *engagement-score* for the entire online activity. To the best of our knowledge, we are the first to envision monitoring engagement by correlating the speaker's intent and listeners' head gesticulations.

(3) **Thorough evaluation and testing:** We evaluate WiFiTuned with 22 participants in 4 different locations with 6 different types of meetings. First, for each participant, we monitor the engagement level in a fine-tuned manner (segment level as shown in Fig. 1) to capture the subjective variations in engagement throughout the meeting. Thus, we classify the segments either as engaged or disengaged. WiFiTuned correctly classifies each segment and achieves an average accuracy and F2_score of 86.82% and 85.70%, respectively. WiFiTuned captures the multitasking behavior of the participants with an average accuracy of 86.405%. WiFiTuned accurately captures the disengagement behavior of the participants in some challenging scenarios, such as participants lying down, changing position frequently, and reading irrelevant papers with an average accuracy of 82.96%. WiFiTuned capture engaged participants with an average accuracy of 91%. This shows that WiFiTuned adapts well across different participants, locations, and meeting types. We compute each participant's engagement score for the entire meeting duration. WiFiTuned computes the engagement score of the participants with an average error

of only 5.86, compared to baseline models with an average error of 12.23. It estimates the engagement score with an average error of 6.2 for the participants involved in multitasking, 5.71 for disengaged participants, and 5.92 for engaged participants.

Notably, an earlier version of this paper was published in [48], where we discussed the core idea of using CSI and Audio information to extract the engagement scores of the participants during a synchronous online meeting. However, the earlier model is restricted to the specific setup of the WiFi transmitter and the receiver and requires controlled placement of the devices to infer the head gesticulations with high accuracy. However, in this extended version of the work, we have developed a novel approach by utilizing the Doppler shift vector from the CSI data to extract the head gesticulations in an environment-neutral way. The thorough evaluation indicates that the enhanced approach improves the system's accuracy significantly (24.31%) compared to the approach presented in the earlier version of the paper under a general in-the-wild setup where the participants do not need to control the placement of the devices. To obtain the system's usability feedback, we deployed and evaluated WiFiTuned in the real scenario with 10 participants independent of those involved in the data collection and collected usability feedback from those participants. Further, we shared the demo video demonstrating the working of WiFiTuned and received online feedback from 33 additional participants. Based on the feedback from those 43 participants, we observe that WiFiTuned attains an average usability score of 79.73% on the system usability scale (SUS), which indicates the overall adaptability of the platform for real-world usage.

2 BACKGROUND AND RELATED WORK

In this section, we first present a background on WiFi sensing. Next, we discuss the literature related to this work.

2.1 WiFi-based Environment Sensing

When a WiFi signal propagates between a transmitter (Tx) and a receiver (Rx), it experiences various impairments such as scattering, reflection, and diffraction. Such impairments are caused by obstacles between the Tx and the Rx, such as walls, furniture, and human subjects. Subsequently, human movements between the Tx and the Rx also impact the signal properties. Notably, the *Channel State Information* (CSI) defines the signal properties captured at the Rx, as defined next.

2.1.1 Channel State Information. The recent WiFi standards like IEEE 802.11ax use multiple antennas using the MIMO (Multiple Input Multiple Output) technology and Orthogonal Frequency Division Multiple Access (OFDM/OFDMA) at the physical layer, which divides the WiFi channel into K orthogonal sub-carriers. The CSI is computed between each pair of Tx and Rx antennas for each sub-carrier $k \in \{1 \dots K\}$ and for each packet $n \in \{1 \dots N\}$. Let H_k^n denote the CSI for packet n received over sub-carrier k, which is computed as follows.

$$H_k^n = A_k^n . e^{j\phi_k^n} \tag{1}$$

Here A_k^n is the attenuation and ϕ_k^n is the phase shift of k^{th} sub-carrier of packet n. Moreover, considering multi-path propagation between the Tx and the Rx, P copies of signals are received at Rx. Hence, H_k^n at Eq. (1) can be written as follows.

$$H_k^n = \sum_{p=1}^{P} A_{k,p}^n . e^{-j2\pi(f_c + k/T)\tau_{k,p}^n} \tag{2}$$

Here, $A^n_{k,p}$ and $\tau^n_{k,p}$ are the attenuation and delay associated with path $p \in \{1 \dots P\}$. f_c is the central frequency of the k^{th} sub-carrier, $T = 1/\Delta f$ is symbol time, and Δf is the sub-carrier spacing.

The WiFi sensing frameworks pre-process CSI values to extract the relevant features. These features are then analyzed to discern changes in the signal properties caused by the humans between the signal propagation path. The feature analysis facilitates learning and recognizing various human movement patterns. The proliferation of WiFi devices provides an opportunity to apply CSI-based WiFi sensing to a broad spectrum of problems, including object localization [14], smart home applications [41], human activity detection [35], as well as several fine-grained applications like hand or finger movement tracking [58], sleep posture monitoring [15], etc. We can extract various features such as amplitude, phase, angle of arrival (AoA), angle of departure (AoD), and Doppler vectors from the CSI values. *Notably, recent studies [9, 46] emphasize that the Doppler vector provides environment-independent features to identify activities across environments.*

2.1.2 Doppler Phase Vector. The Doppler phase vector represents the shift in the signal phase caused by the changes in the geometry of the multi-path propagation during the transmission event caused by the moving scatterer. Let $\tau^n_{k,p}$ at equation 2 be expressed as the sum of two components: $\tau^n_{k,p} = \frac{l^n_{k,p} + \Delta M^n_{k,p}}{c}$, where $l^n_{k,p}$ represents the static component (initial position of scatterer) and $\Delta M^n_{k,p}$ represents the dynamic component (related to scatterer movement) and c is speed of light. Consider a human subject performing some activities in the WiFi signal propagation path. During activities, each body part of a person can be considered a scatterer moving with a speed of $v^n_{k,p}$ related to p^{th} path of sub-carrier k of packet n. Hence, $\Delta M^n_{k,p}$ is defined as $\Delta M^n_{k,p} = -v^n_{k,p} \cos \alpha^n_{k,p}.nT_c$, Where nT_c is the transmission period and $\cos \alpha^n_{k,p}$ is the angle of motion of the moving scatterer associated with the p^{th} path. The activity-related motion of the scatterer with velocity $v^n_{k,p}$ causes a phase shift in the received signal, which, in turn, affects the CSI. This is revealed by the Doppler Phase Vector extracted from the CSI samples through a *Short-Time Fourier Transform* (STFT).

2.2 Related Work

Next, we discuss the prior literature focusing on developing engagement monitoring platforms while utilizing different modalities, such as video, audio, and sensor data.

2.2.1 Vision-based Engagement Monitoring. Vision-based techniques have widely been adopted for estimating overt visual attention, characterized by detectable eye movements and saccades. One of the domains where such vision-based techniques are applied profoundly is the analysis of human engagement in online scenarios by analyzing the participants' visual patterns. For instance, in [29], the authors have utilized the capabilities of a commodity smartphone to track the gazing patterns of learners attending Massive Open Online Courses (MOOC). The system primarily relies on the device's front camera and adopts a vision-based approach to estimate the visual attention level of the learners. Since a learner's engagement or attention is correlated to the different activities performed by them while watching online videos, Zhu *et al.* [61] proposed a novel system that integrates gaze information of online learners with their mouse activities to classify different relevant and irrelevant tasks. Gaze-based engagement estimation has also been extended to physical classrooms and Intelligent Tutoring Systems (ITS). In [26], the authors have presented a novel technique that utilizes gaze-based features like saccades and fixations, extracted by a commercial tracker, to investigate the state of learners' mind wandering while using an ITS.

However, under robust real-world setups, attention can also be covert in nature [25] (paying attention without moving the eyes). Hence, the mere use of gaze information can be limiting and insufficient. Therefore, existing literature

has used expressions [51], facial orientation [47], and related cues as reliable indicators of human engagement [57]. In a study presented in [11], Carolis *et al.* shows that a *Long Short-Term Memory* (LSTM) network, trained with gaze and head orientation features, along with those of a learner's facial *Action Units* (AU) can estimate their engagement in MOOC. Besides these popular techniques for tracking human attention, vision-based techniques have also been applied for extracting physiological data like heart rate [54] that can be associated with various forms of human attention.

Despite their advantages, these vision-based models suffer inherent challenges while coping with different ambient light levels, facial occlusion, free bodily movements, etc. Such drawbacks necessitate the involvement of sophisticated hardware and dedicated sensors for engagement detection. Next, we discuss some prior works proposing sensor-based engagement monitoring.

2.2.2 Sensor-based Engagement Monitoring. Engagement has been measured by unimodal and multimodal sensor data based on the target audience and the availability of resources. In [21], the authors have presented a novel study that aggregates the (dis) engagement level of the audience in a hybrid meeting and conveys it to the presenter through thermal feedback using a wearable device. The study shows how "warmer" feedback can encourage the active speaker in hybrid meetings. Conversely, a "colder" response can lead to a significant distraction, thus impacting the presenter's performance. Similarly, [30] shows how near ultrasound chirp signals can be used for classifying facial expressions that can provide insight into the user's engagement level to online videos.

Multimodal Engagement Monitoring. Apart from these unimodal approaches, [16] adopts a multimodal approach by employing commercial smartwatches to capture physiological data like electrodermal activity (EDA) and heart rate, along with accelerometer reading of the user. These learner-specific readings are integrated with environmental data like CO2 level, humidity, temperature, and sound level of the classroom to predict the learners' engagement level. These modalities have also been explored for understanding the engagement of autistic children in human-robot interactions [44]. Several other works have integrated the users' various visual and physiological features to understand their state of engagement in various scenarios [27, 37].

Although dedicated sensor-based techniques have successfully overcome the challenges of vision-based techniques, they are often intrusive and costly. In this paper, we explore non-intrusive passive modalities like WiFi and audio that can be sensed pervasively on the meeting devices and develop a non-vision-based technique for continuous attention monitoring during synchronous online meetings. We next discuss a set of surveys and pilot studies to highlight and motivate the problem space.

3 MOTIVATION

The core idea of our model is that the overall body language and head gestures of the attentive participants during a synchronous online meeting should corroborate with the meeting's discussions. Consequently, the problem of continuous attention monitoring can be divided into three sub-problems: (1) understand the head gesticulation patterns of the participants during the online meeting, (2) extract the intents of the discussion captured from the meeting audio, and (3) correlate the head gestures with the discussion intents captured over time. To show that these hypotheses hold for general synchronous online meetings, we conducted a large-scale anonymous survey to study collective and common head gesticulation patterns related to engagement levels of users in online meetings. A total of 334 participants (73.7% male and 23.4% female) above 18 years of age from India, Kuwait, the United States, and the United Kingdom responded to the survey. 85.9% were students, 7.8% were scholars, and 3% belonged to the academia & IT industry.

45.2% participants attend online meetings more than once a day, 35.1% once a day, and 19.7% once a month. Notably, we observed that a laptop (97.3%) is the most preferred device for such activities.

3.1 Scope and Nature of Multitasking

We asked the participants to select multiple types of activities that they perform while attending online meetings. There were three main hypotheses in this regard. *(a) The majority of the participants multitask during online meetings:* 91.30% participant multitask during online meetings, while only 8.7% concentrate on the meeting without multitasking. *(b) The participant performs various types of multitasking:* The participant involves in *"taking notes" (*78.7%*)*, *"checking emails, playing games, reading/sending texts, etc. on smartphones"* (65.3%), *"open a different tab in the browser and check websites/emails"* (67.1%), *"eating food"* (53.9%), *"watching other videos"* (24.6%), *'dozing"* (10.3%), and others (2.4%). (c) *Due to the lack of supervision, some parallel activities contribute to disengagement:* 62.45% of the participants perform certain tasks due to lack of supervision that contribute to disengagement, such as "eating food", "using smartphones", "checking emails", "watching videos", "dozing", etc.

3.2 Head Nod/Shake and Engagement

We asked the participants if they nod/shake their heads while engaged. We hypothesize that *head nodding/shaking should indicate engagement and will not depend on whether the participants are visible to each other.* To this, 88.9% of the participants mentioned that they would nod/shake while engaged (actively listening and agreeing/disagreeing with the discussion). Further, 93.4% engaged participants would nod/shake when visible, and 82.6% engaged participants would nod/shake when not visible. Hence, nodding/shaking is an indication of engagement, and it does not depend on whether or not the participants can see each other in meetings.

3.3 Gaze Gesticulations and Engagement

We aim to understand whether gazing behaviour can be associated with engagement. We hypothesize that *different gaze (hence head) gesticulations can be associated with the (dis)engagement of the users.* 91.9% of the respondents mentioned that gazing at the screen and listening to the speaker would imply engagement. 80.8% agreed that looking around would imply distraction and, hence, disengagement. For most participants, shorter gaze downs are caused by engagement-enhancing activities like taking notes and hence would imply engagement (68.8%), and longer gaze downs would imply disengagement (44.9%).

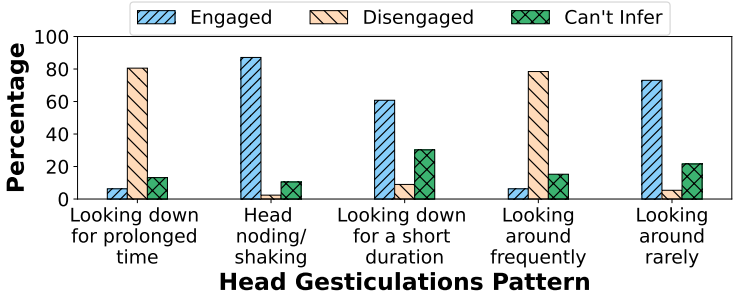


Fig. 2. Mapping between different head gesticulation patterns and engagement indicated by the participants.

Further, we asked the participants to provide a rating between 1 – 5 (1 – *do not reveal engagement*, 5 – *completely reveal engagement*) to the head gesticulations pattern. 88.7% participants mostly/always (4 – 5) find head gesticulations highly helpful in indicating the engagement level. Fig. 2 summarizes how different types of head gesticulation patterns indicate engagement or disengagement.

The key takeaways from this survey are as follows: **Firstly**, multitasking is an integral part of online meetings, and some of these activities lead to disengagement. Thus, it becomes crucial to monitor a person's level of engagement. **Secondly**, many parallel activities during online meetings involve visual context switching away from the primary monitor. Therefore, a new modality is necessary to monitor engagement levels accurately. **Thirdly**, head gesticulations of engaged participants are correlated with the speaker's intent. **Finally**, the survey reveals that head gesticulation is a reliable indicator of (dis)engagement and is invariant to the visibility of the meeting participants to each other. Thus, the head gesticulation-based system monitors engagement without the need to open the camera. We utilize these takeaways when designing our system.

4 WIFITUNED DESIGN

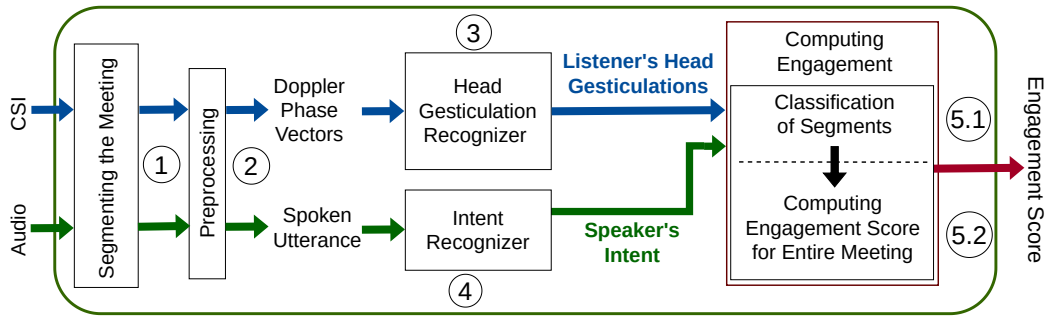


Fig. 3. Design of WiFiTuned: *Input: WiFi CSI & Audio. Output: Engagement score.* Our system is novel in considering the way of estimating the engagement score by correlating the speaker's intent and the listener's head gesticulations.

Fig. 3 shows the design overview of WiFiTuned. It takes *WiFi CSI* and *meeting audio* as input and provides an *engagement-score* as output for each participant. It has five modules: (1) *Segmenting the meeting*: as multitasking and distractions are common in meetings, we divide the entire meeting duration into smaller segments to monitor the engagement in a fine-tuned manner; (2) *Pre-processing*: WiFi CSI is impacted by the environmental clutter, interferers (persons moving around), and static objects. Hence, we pre-process the CSI data to obtain environment-independent features called Doppler vectors. Similarly, we pre-process the audio data to get spoken utterances; (3) *Head gesticulation recognizer*: a robust model to recognize the head gesticulations in different environmental settings; (4) *Intent recognizer*: a model to recognize the speaker's intent; and (5) *Computing engagement*: first, it classifies each segment as either engaged or disengaged depending on whether head gesticulations are in tune with the audio intent. Then, it computes the engagement score for the entire meeting. The detail follows.

4.1 Segmenting the Meeting

The large-scale survey indicates that multitasking and momentary distractions are common in online meetings. Moreover, participants might not be engaged continuously for the entire duration of the online meeting. For example, a participant

might appear engaged early but disengaged at the end. Considering multitasking, momentary distraction, and varied engagement duration, it is essential to monitor the engagement in a fine-tuned manner to capture the variations in engagement throughout the meeting. Hence, we divide the entire meeting duration into smaller T second segments. For each segment, we pre-process the CSI and audio data to compute the Doppler vector traces and spoken utterances, respectively.

4.2 Pre-processing

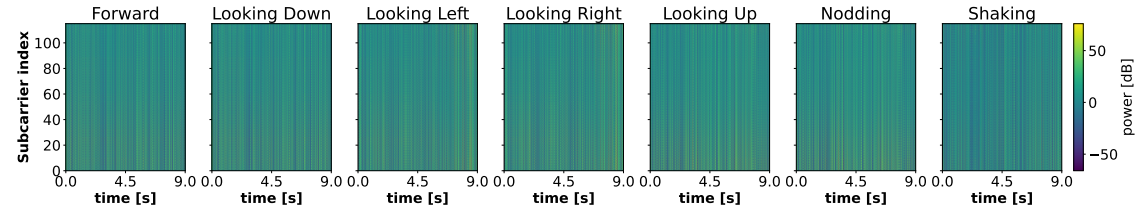
We aim to monitor and recognize the head gesticulations of the participants in various indoor locations with different environmental settings. We utilized CSI of WiFi signals for the same. However, as shown in Fig. 4a, directly utilizing raw CSI data does not allow robust activity recognition. The CSI is affected by surrounding static objects, interferer (persons moving around), and environmental settings. The noise significantly obscures the delicate multi-scale variations induced by the head gesticulations in the CSI data. Next, we extracted amplitude from raw CSI data like prior works [2, 7, 59]. However, as shown in Fig. 4b, even such existing methods fail to capture the minute multi-scale variations. Next, we followed a phase sanitization procedure like in prior work [35]. It removes the undesired phase offset from the phase traces shown in Fig. 4c. However, due to the environment-dependent nature of CSI, the sanitized phase does not provide environment-independent features (see the temporal patterns of the head gesticulations in Fig. 4c). This impacts a recognition model's ability to train effectively. Consequently, the recognition model fails to capture the most relevant features from the input data to recognize head gesticulations. Such environment-dependent nature of CSI data makes head gesticulation recognition challenging, as environmental variations hinder recognizing correct head gesticulations. We thus exploit the Doppler effect to obtain environment-independent features to recognize head gesticulation. For this, we pre-process the CSI data to remove the environmental noise and compute the *Doppler phase vectors, which provide the Doppler effect of each head gesticulations* (details in §.2.1.2). Fig. 4d shows the Doppler vector traces computed from sanitized CSI data. The Doppler vector traces show the variation in the Doppler effect changes in response to different head gesticulations. The temporal patterns exhibit different head gesticulation patterns compared to raw CSI, amplitude, and sanitized phase. However, such a pattern is not very pronounced with the presence of environmental-induced noise. We next remove the environmental noise and compute the Doppler traces shown in Fig 4e, which shows the precise Doppler effect of different head gesticulations. The denoised CSI provides accurate Doppler phase vectors. The Doppler vector traces generated by merely sanitizing the phase do not allow environmental noise to be removed. Notably, We labeled the CSI data by utilizing the participant's video feed. We pre-process the meeting audio and extract the spoken utterance from the meeting audio to recognize the speaker's intent. The details follow.

4.2.1 Pre-processing of WiFi CSI and Doppler Phase Vector Computation. We first apply denoise procedures to obtain clean CSI data. Next, we compute the Doppler phase vectors.

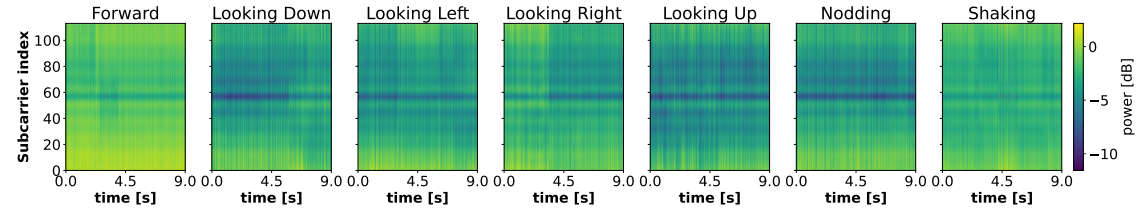
Denoising CSI: The environment's static/dynamic objects (S, D) and interferer (I) induce noise. Hence, the estimated CSI H_k^n can be defined as:

$$H_k^n = H_{k_{(T)}}^n + H_{k_{(S,D)}}^n + H_{k_{(I)}}^n \tag{3}$$

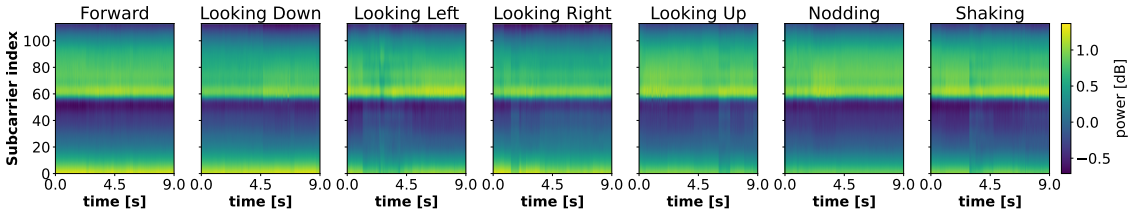
Here $H_{k_{(T)}}^n$ is the CSI from Tx to Rx impacted by the target person T, $H_{k_{(S,D)}}^n$ and $H_{k_{(I)}}^n$ is the CSI (considered noise component) from Tx to Rx impacted by static/dynamic objects (S, D) and interferer I respectively at receiver of sub-carrier $k = \{1 \dots K\}$ for packet $n = \{1 \dots N\}$. Moreover, the hardware artifacts induce phase offsets (changes in the phase values), such as carrier frequency offset (CFO) and sampling frequency offset (SFO). The H_k^n with the undesired



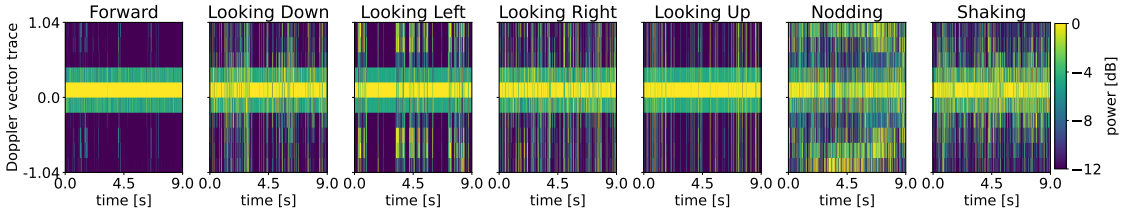
(a) Raw CSI traces.



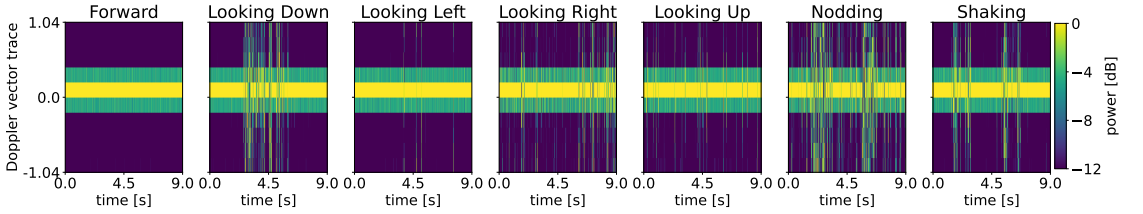
(b) Amplitude traces computed from raw CSI data.



(c) Phase traces computed from sanitized CSI data.



(d) Doppler vector traces computed from the sanitized CSI data.



(e) Doppler vector traces computed from the sanitized and denoised CSI data.

Fig. 4. Each trace is 9 second long and shows the variation of each head gesticulation.

phase offset can be defined as:

$$\bar{H}_k^n = H_k^n e^{\bar{\phi}_k^n}. \tag{4}$$

Here $\bar{\phi}_k^n$ is undesired phase offset and $\bar{H}_k^n$ is the resultant CSI of sub-carrier k for packet n. Phase offset induces distortion in the signal pattern associated with activities, leading to misclassification. Moreover, the recognition model trained on CSI data with phase offset lacks robustness when deployed in a new environment with different phase offset noise. Thus, we apply the sanitization procedure [35] to remove the undesired phase offset (SFO and CFO) and retrieve H_k^n. Next, we deployed denoising techniques explained in [49] to remove the noise component from the estimated CSI to preserve the component ($H_{k_{(T)}}^n$) affected by the target person's (T) head gesticulation. Firstly, we apply a Hampel filter [39] to remove the high-frequency noise/anomalies caused by the environmental factors and hardware artifacts. Secondly, we apply a 1-D Wavelet transform (DWT) filter [52] to remove noise caused by surrounding objects and people moving around. Finally, we use the Savitzky-Golay smoothing filter [34] to preserve the fluctuation induced by head gesticulations. After phase offset removal and denoising procedure, the computed CSI ($H_{k_{(T)}}^n$) data samples are used to compute the Doppler vectors.

Doppler Phase Vector computation: Let H_K^N where $H_{k_{(T)}}^n \in H_{K(T)}^N$ denote the CSI dataset, where K is the total number of sub-carriers, and N is the total collected packets as:

$$H_K^N = \begin{bmatrix} H_1^1, \ldots, H_1^N \\ \vdots \\ H_K^1, \ldots, H_K^N \end{bmatrix} \tag{5}$$

To compute the Doppler phase vector, we define the observation window $i \in \{1 \ldots W\}$ that specifies the number of packets $R < N$ to estimate the changes in path delay (phase shift). The total number of observation windows is $W = N - (R - 1)$. We compute the Doppler velocity $d_{v,i}$ for i^{th} observation window using $H_{K(T),i}^R$ as:

$$d_{v,i} = \sum_{k=1}^{K} |\mathcal{F}(H_{K(T),i}^R|^2, \tag{6}$$

where $\mathcal{F}$ represents the Short-Time Fourier Transform (STFT), and the absolute values of the square of the resulting STFT coefficients are summed over all sub-carriers k. The collection of V number of Doppler velocities forms a Doppler Phase Vector D_i as:

$$D_i = [d_{1,i} \ldots d_{V,i}]^T \tag{7}$$

We compute the next Doppler Phase Vector by sliding the observation window with $strides = 1$ and so on. *In the following, we compute the Doppler Phase Vector traces (dimension $W \times V$) by stacking subsequent Doppler Phase Vectors, where each row represents the Doppler Phase Vector D_i for observation window i.* The labeled Doppler vector traces are used to train the head gesticulation recognition model.

4.2.2 Speaker Spoken Utterances. A spoken utterance is a segment of speech spoken by the speaker. Each spoken utterance is a distinct unit that conveys meaning and intention. For example, the segment of speech "*HCI aims to improve the interactions between users and computers by making computers more usable and receptive to the user's needs*," is an example of spoken utterance that conveys "*meaning*" and "*explanation*" intent. Hence, we utilize the spoken utterances to recognize the speaker's intent for each segment. We extract the spoken utterance in text using *Google Speech Recognition* service `recognize_google` for each segment.

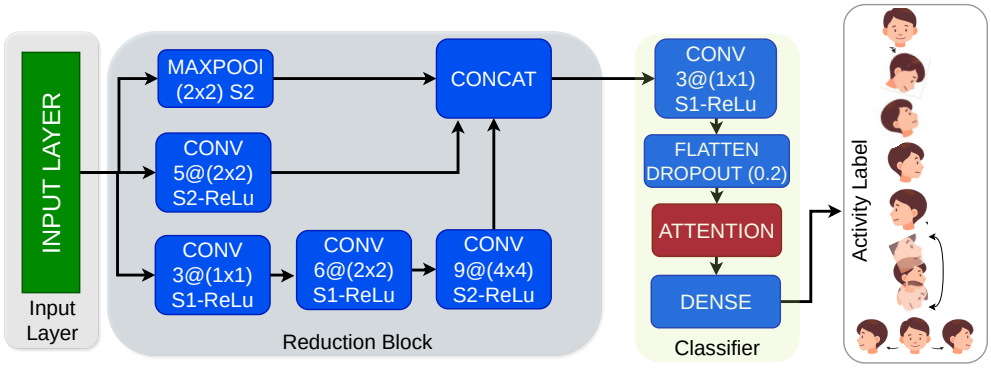


Fig. 5. WiFiTuned's head gesticulations recognition model: SHARPA.

4.3 Head Gesticulation Recognizer

The Doppler phase vector captures multi-scale variations in the WiFi CSI induced by the user's head gesticulations. A head gesticulation-recognizing model must learn these variations to recognize different head gesticulations. Thus, we utilize *state-of-the-art SHARP (Sensing Human Activities through WiFi Radio Propagation) [35]* neural-network model for head gesticulation recognition. The architecture of the model is shown in Fig. 5. The model takes the Doppler phase vector traces as input and returns the label of the head gesticulations being performed. First, the input is fed to the *Input Layer* connected to the *Reduction Block*. Each head gesticulation affects the Doppler phase vectors at varying scales; for example, nodding and looking forward cause different scale variations in the Doppler phase vectors. Hence, the Reduction Block extracts the most significant features from the Doppler phase vector traces at several scales to capture these variations by using *MAXPOOL* and convolutional (*CONV*) layers in a parallel fashion. The *MAXPOOL* layer reduces the input dimension and retains the most significant features in the input. The *CONV* layers process the input features through different-sized kernels (such as 2×2, 1×1, and 4×4) with strides $S1 - 1 \times 1$ and $S2 - 2 \times 2$, followed by *ReLu* activation function. The strides represent the movement of the kernels horizontally and vertically. Each kernel processes the input features and extracts features (called feature maps) at a specific scale, depending on its size. Next, the ReLu activation function is applied to the feature maps to learn more complex patterns. Thus, the Reduction Block obtains the multi-scale features that better represent the input. The feature maps obtained from the *MAXPOOL* and *CONV* layers are concatenated (*CONCAT*) and passed to *Classifier*. The Classifier consists of *CONV*, *FLATTEN*, *DROPOUT*, and *DENSE* layers. The output of the CONCAT layer is passed to a CONV layer with a 1×1 sized kernel to reduce the feature maps. Next, the output of that layer is flattened using the FLATTEN layer and passed to the DROPOUT layer. The DROPOUT layer applies a dropout regularization technique on the flatted feature vector. The dropout regulariser randomly zeroes 20% elements in the flatted feature vector preceding the DENSE layer with 7 output neurons, one for each head gesticulation.

The SHARP model is claimed to be environment-independent and person-independent. However, we observe that it could not adapt to environmental dynamics induced by in-the-wild settings where the placement of Tx and Rx is not controlled (details in §. 6.1.2). The CSI data is generally noisy, but the data collected in real-world settings is particularly challenging due to increased noise levels. As a result, Doppler phase vector traces contain more irrelevant and unstable input features. The model often learns irrelevant and unstable features and thus becomes prone to overfitting. Moreover,

the recognition model must focus on the most relevant input features and ignore suppressing noise & irrelevant features. The model must prioritize the most stable input features and discard unstable input features. Consequently, we incorporate an Attention (*ATTENTION*) layer of 64 nodes after the DROPOUT layer, as shown in Fig. 5. We select the attention nodes empirically to ensure an optimal trade-off between model performance and computational efficiency. The ATTENTION layer enables selective focus on the input features to generate output. It scores the feature vector to prioritize the most stable input features. Hence, the model can better learn the complex multi-scale variations to make accurate predictions and make the recognition model robust across environmental dynamics, interference, and random noise. The output of the ATTENTION is connected to a dense layer of seven neurons, followed by a softmax to produce head gesticulation classification results. The SHARPA model is trained in a supervised manner on labeled Doppler vector traces using the categorical cross-entropy loss function.

4.4 Intent Recognizer

We use a pre-trained intent recognition model *klue/roberta-small* [38] trained on 3*i*4*k* dataset with 55134 samples to recognize the speaker's intention from spoken utterances. The model returns seven intention labels such as *fragment*, *statement*, *questioning*, *command*, *rhetorical question*, *rhetorical command*, and *intonation-dependent utterance*. We merge labels as per the speaker's context. For example, a participant's head gesticulations might be similar if it is a *questioning* or *rhetorical question*. Hence, we group questioning & rhetorical question as *questioning* and command & rhetorical command as *command*. Since statements (and related intents like fragments and intonations) use declarative clauses that are more aligned with informative speeches, they are grouped under the *explaining* label [42].

Table 1. Specific behaviours of (dis)engaged listeners inferred from their head gesticulations tuned with speaker's intent.

Label	Intent	Head Gesticulations	Specific Behaviours
Engage	Explaining	Looking Forward	Looking at the screen and following the lecture content
		Looking Down	Writing notes, looking down for a short duration
		Looking Down, Up, Right, and Left	Looking around less frequently and for short duration
Engage	Questioning	Nodding/Shaking	Giving feedback, agreeing to a discussion
		Looking Forward	Listening to the speaker, looking at the screen.
		Looking Down, Left, Right and Up	Looking around for short duration and less frequently
Engage	Command	Nodding/Shaking	Respond to the speaker
		Looking Forward	Reading notes
		Looking Down	Taking notes, solving relevant questions
		Looking Up, Right, Left	Looking around for short duration and less frequently
Disengage	Explaining	Looking Up, Down, Left and Right	Looking around for a long duration and more frequently
	Questioning	Looking Up, Down, Left, Right	No nodding and shaking
	Command	Looking Down/Up	Looking down or looking up for the long duration

4.5 Computing Engagement

From the above modules, we finally obtain the head gesticulations and audio intent, which works as an input to computing engagement as shown in Fig. 3. We monitor the engagement level of the participants at the segment level. In each segment, we analyze the head gesticulations of the participants to infer their engagement level, which can be revealed by quantifying the expected and unexpected head gesticulations. Table 1 shows the expected and unexpected head gesticulations with respect to each intent based on the takeaways from our survey. The expected head gesticulations are in tune with the speaker's intent and are indicators of engagement. For example, engaged participants nod or shake their heads in response to giving feedback or agreeing in a discussion, aligning their gestures with the speaker's intent, such as when the speaker is questioning. The head gesticulations that are not in tune with the speaker's intent are considered unexpected and characterize disengagement. For example, disengaged participants frequently look around (not actively listening) when the speaker is questioning. However, unexpected head gesticulations for a short duration are part of positive multitasking and indicate engagement, as shown in Fig 2 (§.3). We utilize Table 1 to quantify the engagement level of the participants by classifying them as either *engaged* or *disengaged*.

A simple mechanism for classifying segments based on the expected and unexpected head gesticulations can be as follows: one can assign +1 to expected head gesticulations and −1 to unexpected head gesticulations based on Table. 1. Next, for each segment i, one can compute the *engagement-metric* $E_i = (\sum((+1 \times f_{en}) + (-1 \times f_{dis})))/f$. Where f_{en} is the total frequency of "expected" head gesticulations, f_{dis} is the total frequency of "unexpected" head gesticulations, respectively, and f is the total number of head gesticulations. One can then classify each segment as engaged or not based on whether *engagement-metric* > a threshold **d**. *However, determining an optimal threshold for each individual segment and for each person is challenging as engagement is subjective in nature and varies from person to person.* The frequency of each head gesticulation with respect to the same speaker's intent might differ due to environmental distractions, meeting content, and individual cognitive processes. For example, in a segment, one person nods frequently to express agreement and understanding, while another looks forward and down frequently to reflect on the information and take notes. Here, both are actively engaged, but they seem to be different. Similarly, in certain segments, individuals nod and look around frequently in multiple short durations. This behaviour indicates varying levels of engagement. Moreover, an individual's head gesticulations vary over segments for the same intent, corresponding to different levels of engagement. Thus, some or all segments may be classified as engaged or disengaged. These examples illustrate how diverse the head gesticulation patterns might be, making it challenging to apply single threshold **d** uniformly across segments and individuals for efficient classification of segments using engagement metrics. *This insight obliges us to analyze how frequently head gesticulations occur, the intricate patterns the head gesticulations form, and whether frequencies and patterns align with the underlying intent.*

To address the above challenges, we process the segments and group them into two hierarchies: first based on the head gesticulation pattern and then on the speaker's intent. Grouping the segments based on head gesticulations that follow a similar pattern and exhibit partial similarities in frequencies. Hence, grouping the segments with comparable head gesticulations reduces the impact of frequency difference. We consider each segment within a group to have the same level of engagement irrespective of different frequencies of head gesticulations. Grouping the segments within a group into sub-groups allows the model to efficiently contextualize different head gesticulation patterns with respect to the speaker's intent. For example, in a group, segments consist *looking forward, followed by looking down* head gesticulations pattern might indicate the intent type of *explaining, questioning and command*. Looking down during a command and explaining relates to engagement, such as taking notes or being thoughtful. In contrast, looking

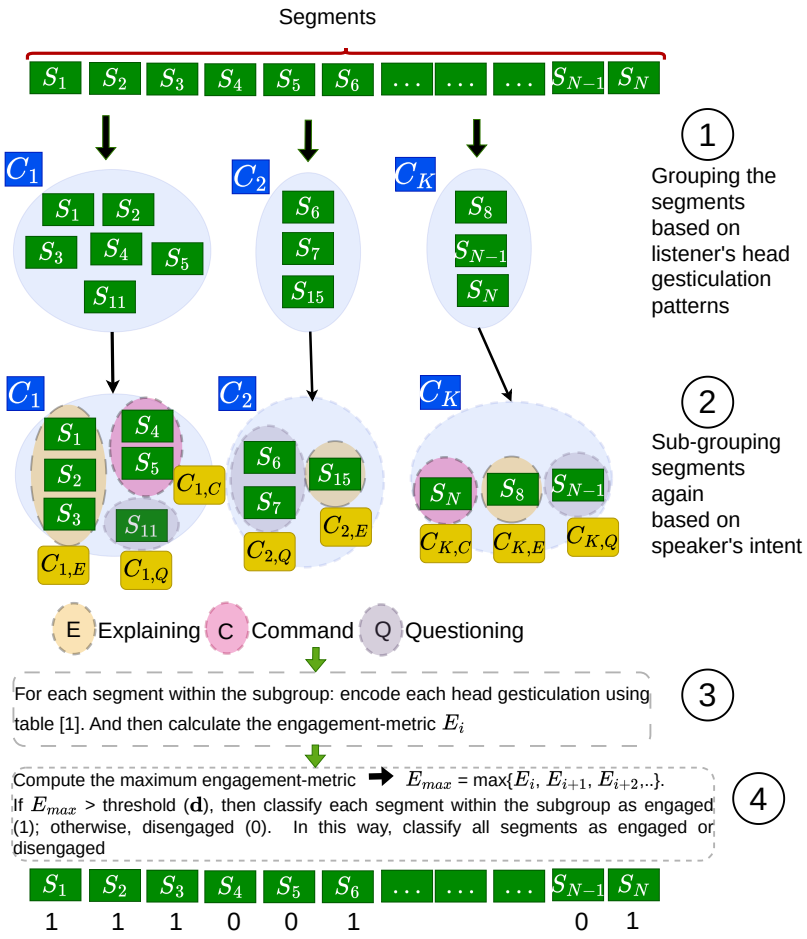


Fig. 6. Classification of segments as engaged(1)/disengaged(0).

down while questioning may be related to momentary distraction. Only after grouping the head gesticulations once by patterns and next by intent, we arrive at a point where specific intent-wise head gesticulation patterns are categorized into a single group. At this point, we utilize our Table. 1 to encode the head gesticulations and compute the engagement metric E_i for each segment with the subgroup. Given that the segments within the subgroup are similar in both terms of patterns and intent, we can now utilize a single threshold to classify all the segments with that sub-group as engaged or not. Thus, we address the challenge of selecting different thresholds for different segments. Next, we provide the details of classifying segments and computing engagement scores.

4.5.1 Classification of segments. Fig. 6 shows the overall classification process. The meeting is divided into smaller segments denoted as $S_1, S_2, \ldots, S_N$, where N is the total number of segments. The process of classifying segments involves 4 steps as follows. ① We create groups of segments denoted as $C_1, \ldots, C_K$ based on the similarity of head

gesticulation patterns, where K is the total number of created groups. We use *Gestalt Pattern Matching algorithm [28]* to estimate the similarity score. The Gestalt algorithm compares the patterns of head gesticulations and obtains a similarity score for each pair of segments. This similarity score reflects the degree of temporal resemblance between the patterns. For example, if a participant looks forward in two segments, the Gestalt algorithm obtains a high similarity score. However, if a participant looks forward in one segment and looks around in another, the similarity score obtained by the Gestalt algorithm is low. We apply *threshold-based clustering* strategy to create the groups of segments. The segments with similar head gesticulation patterns (similarity score $\geq$ 70%) are grouped together. We do not fix the number of groups, which may vary from meeting to meeting. ② Each group contains segments that exhibit various intents. For example, within a single group C_2, some segments (S_6 and S_7) may exhibit a *questioning* intent, while others (S_{15}) may exhibit an *explaining* intent. Hence, we further sub-group segments within each group based on intent. The sub-grouping of segments, based on intent, enables the analysis of diverse head gesticulation patterns across intents within each group. ③ We encode head gesticulations in each segment, considering the speaker's intent. We assign +1 to expected head gesticulations and -1 to unexpected head gesticulations. In the case of looking down, we assign 0 if it belongs to explaining or commanding and -1 if it belongs to questioning. Looking down conveys different behaviour depending on the intent context. Next, for each segment within the sub-group, we compute the *engagement-metric* E_i. ④ Next, we compute the maximum engagement-metric (E_{max}) in that sub-group. Considering E_{max} to classify each segment within a subgroup eliminates the impact of the frequency difference of head gesticulations. Each segment in a sub-group is classified using E_{max}. Thus, each sub-group and its associated segments are classified as engaged (1) if $E_{max} > \mathbf{d}$; otherwise, it is classified as disengaged 0. The classification of segments into two categories, engaged and disengaged, is considered a binary classification. Thus, we empirically set $d = 0.5$. We repeat the process for each sub-group. Finally, we obtain the engagement label (either $1 - engaged$ or $0 - disengaged$) of each segment. Next, we compute the engagement score of the entire online activity.

4.5.2 Computation of engagement score of entire online participation. After classifying the segments as engaged or disengaged, we scan each segment one by one sequentially. We increment the value of engagement score *En* by 1 if the segment is classified as *engaged*. However, if any continuous segment is classified as *disengaged*, we do not decrement *En* immediately. Intuitively, a user might not be engaged continuously for the entire duration of the online meeting. A momentary distraction for a short duration does not indicate disengagement, but being distracted for longer might do. Hence, we check the next n segments. If the continuous segments are classified as disengaged, we decrement *En* by the total number of *disengaged* segments, otherwise, ignore the disengaged segments. We empirically find the optimal value of $n = 18$. The final engagement score is computed as $E = En/S_N$, where S_N is the total number of segments.

5 EXPERIMENT METHODOLOGY

To evaluate WiFiTuned, we ask the following research questions to analyze its performance under different aspects.
RQ1: How well can WiFiTuned recognize the participant's head gesticulations using CSI across diverse deployments (whether semi-controlled or uncontrolled in-the-wild?) We hypothesize that the head gesticulation recognizer accurately recognizes each head gesticulation both in semi-controlled and in-the-wild deployments.
RQ2: To what extent can WiFiTuned accurately classify segments as engage or disengage in various deployment scenarios, ranging from semi-controlled environments to uncontrolled in-the-wild settings? We hypothesize that WiFiTuned classifies each segment correctly, considering each participant behaves differently in online meetings. We further hypothesize that WiFiTuned correctly computes the engagement score of each participant.

RQ3: How well can WiFiTuned perform with different meeting contents and in various locations? We hypothesize that WiFiTuned performs well with varying meeting contents and for multiple locations.

5.1 Evaluation Methodology

5.1.1 Participants Details. We recruited 25 participants, 20 – 30 years of age (10 female and 15 male), who attend online meetings frequently (once a day). The participants were well-informed about the tenure of the study, the seating arrangement, and the study location. Besides this, they were informed that the sessions were being recorded, including their facial previews, for the purpose of research. We shared a consent form regarding the usage of their recorded videos for this study, approved by our Institute's IRB. The objective of the study was not revealed to them in order to avoid performance bias. They were instructed to maintain their natural behaviour during the meeting and had the freedom to be engaged or disengaged as per their interest in the meeting's content. Each participant was compensated with a 6.12 *USD* Amazon coupon.

5.1.2 Meeting Content Types: We scheduled six online meetings in three different sessions (max 1 hour/session). Every meeting has only one participant at a time. We selected six pre-recorded meetings (selected based on popularity) with presentation slides to present in meetings. The meeting contents are different in type and length. Hence, we categorize the meeting as follows– **M1:** *questionnaire* type & length *15 min*, **M2:** *lecture* type & length *45 min*, **M3:** *programming workshop* type & length *17 min*, **M4:** *short lecture* type & length *7 min*, **M5:** *tutorial* type & length *25 min*, and **M6:** *lecture* type & length *16 min*.

5.1.3 Locations: The experiments were conducted in four different indoor environments/locations chosen on the basis of participants' preferences. **Meeting room** (empty room), **Lab-1** (empty lab), **Hostel room** (shared room with another person), and **Lab-2** (crowded research lab). In the meeting room and Lab-1, other people were not present during the experiment. In the hostel room and Lab-2, other people were present. Each participant joins from his/her own preferred location. 6 participants joined from the meeting room, 10 from Lab-1, 3 from the hostel room, and 3 from the Lab-2.

5.1.4 Experiment Setup: All participants join the online meetings using the *Zoom* platform with WiFiTuned setup. The setup has a laptop (with Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz with 8 GB RAM) with an embedded microphone and webcam and two WiFi-enabled ESP32 microcontrollers (transmitter (Tx) and receiver (Rx) with single antenna) as shown in Fig. 1. Both Tx and Rx are flashed with CSI Tool Kit [23] that directly provides raw CSI data. Tx and Rx support the IEEE802.11n standard with a 2.4 GHz band (40 MHz channel with 108 data sub-carriers). Tx is analogous to a WiFi router and is connected to a power source. Rx is connected to the laptop to collect the CSI data. The frequency of data collection is 100 samples per second. Meeting audio and video feeds are recorded from the laptop's mic and webcam.

Semi-controlled setup: The participant sits between LOS of Tx and Rx (placed 1 – 2 meters apart) close to Rx.
In-the-wild setup: In real scenarios, users join online meetings in diverse environment settings that include obstacles between the Tx and Rx, varying distances between Tx and Rx, wall-mounted Tx, varying angles between Tx and Rx, the presence of interferer near the user, and crowded environments.

5.1.5 Segment duration. To monitor the engagement in a fine-tuned manner, we divide the meeting segment into smaller Ts second segments. Whitehill et.al [51] observed that 10s segment duration is reliable and more intuitive for monitoring engagement and ground truth annotation. Moreover, some prior work [13, 29, 50] used similar segment duration (10s). Similarly, we set the $T = 10s$.

5.1.6 Labelling CSI data. We utilize the participant's video feeds to label the raw CSI data with head gesticulations listed in Table 1. We implement the codebase using MediaPipe [1] and PnP algorithm [40] to recognize head gesticulations from the video feed. MediaPipe detects the face from a video feed and feeds a neural network on the detected face to determine the 468 facial landmark. Next, we extract the relevant landmarks (28), such as the corner of the eyes, nose tip, forehead, and mouth corners. Next, the PnP algorithm uses these landmarks to detect the head pose based on its projection to the camera coordinate system in 3D space. After that, 5 head orientations for each frame, such as *looking up, down, forward, left, and right*, are determined. Thereafter, the temporal relation among head orientations is scanned to recognize head gesticulations across 30 frames in a second. For example, nodding involves continuous movement of the head up and down vertically, comprising looking up (down), forward, and down (up) orientations. Head orientations are timestamped. Thus, each head gesticulation corresponds to a 1sinterval. Moreover, each CSI sample is associated with a timestamp. By matching these timestamps, we label each 1s CSI sample with the corresponding head gesticulation.

Table 2. Dataset Description

			Head Gesticulation label				
Setting	Description	Total samples	**Forward%**	**Looking Down%**	**Looking Left/right%**	**Looking Up%**	**Nodding and Shaking%**
Semi-controlled	Tx and Rx placed $1-2$ meters apart	3416708	43.47%	15.37%	13.4%	10.12%	18.00%
In-the-wild	Increasing distance $(1-5)$ between Tx and Rx	138893	25.22%	17.83%	7.14%	9.56%	16.53%
	Tx wall mounted	162495	19.49%	12.27%	8.86%	9.59%	20.44%
	Changing angle between Tx and Rx	114687	20.46%	14.64%	12.38%	6.29%	16.91%
	Obstacle between Tx and Rx	151421	29.52%	23.56%	7.045%	4.23%	14.28%
	Interferer near to target user	73107	7.32%	11.50%	13.87%	9.59%	22.7%
	Tx and Rx in different locations	111704	16.53%	23.80%	10.11%	3.21%	15.47%
	Crowded environment	189699	16.22%	20.253%	1097%	7.28%	17.13%

5.2 Dataset Description

The details of the CSI datasets are listed in Table 2.

Semi-controlled setup: We collected a total of 3416708 time-stamped CSI samples from 132 meetings with a total duration of 42 hours and 35 minutes. The average number of collected CSI data samples for each participant is 155300. We collected a total of 132 (15246 10s segments) meeting audio and frontal video feeds.

In-the-wild setup: We collected a total of 942006 time-stamped CSI samples from 30 sessions with a total duration of 2 hours and 40 minutes.

5.3 Ground Truth Generation

We recruit 15 independent annotators to manually annotate the video segments as *engage* or *disengage*. This manual annotation serves as the ground truth for *segment classification*. We split the frontal video feed of each participant into Ts video segments. We developed a website for annotation where each annotator observes the participant's behaviour in the video segments and annotates accordingly. The video segments are randomly allocated to the annotators. Thus, no single annotator repeatedly annotates the same type of behaviour or participant, reducing personal bias. Three

independent annotators annotate each video segment. Whenever there is a disagreement (around 37% of the cases in which the engage samples are 56% and the disengage samples are 44%) among the annotators, we use a majority vote to mark a video segment as *engage* or *disengage*. We have 67.96% engage label and 32.04% disengage label.

5.4 Baselines

We compare the performance of WiFiTuned with *Gestatten* (gaze gesture-based model) [29]. *Gestatten* extracts the centroid of the user's left and right iris through binarization and correlates it with the movements of prime objects of focus in the meeting, such as instructors, background texts, etc. By correlating the gaze gesture with that of the prime object's trajectory, *Gestatten* generates an engagement score. We also compare WiFiTuned with a simple *Rule-based* baseline– a naive solution for computing engagement score. It finds out the most occurred head gesticulation in a segment and obtains intent from audio data. Next, it classifies each segment as *engage* or *disengage* based upon whether the most occurred head gesticulation is "expected" with respect to the intent (Table 1). Finally, it computes the engagement score by computing the ratio between engaged and total segments.

6 EVALUATION RESULTS

In this section, we evaluate the performance of WiFiTuned and compare it with ground truth and baseline models. For **RQ1**, we evaluate how accurately WiFiTuned recognizes head gesticulations in different environment settings. For **RQ2** and **RQ3**, we evaluate how accurately WiFiTuned classifies each meeting segment as engaged or disengaged and how accurately it computes engagement score of participants?.

Table 3. Performance of recognition models in semi-controlled settings

Model →	**SHARPA**	SHARP	Bi-LSTMA	Contrastive	XGB	Constructive	SVM	GB	RF	LR
Accuracy	**98.47%**	95.00%	94.00%	92.34%	92.23%	57.25%	86.00%	60.00%	53.00%	57.00%
F1-score	**97.71%**	94.76%	93.57%	87.92%	93.50%	57.25%	85.17%	52.35%	25.14%	32.14%
Mean Acc.	**98.3%**	93.00%	94.41%	92.23%	63.17%	52.35%	60.37%	51.45%	35.50%	34.26%

6.1 RQ1: Head Gesticulations Recognition Performance

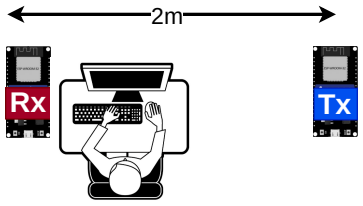


Fig. 7. Semi controlled setup in Lab-1, Lab-2, meeting room and hostel room.

We modeled the head gesticulation recognition as a multi-class classification problem. WiFiTuned takes Doppler vector traces computed from collected CSI data and deploys SHARPA to recognize the head gesticulations. We compute the *Doppler vector traces* by setting $R = 64$ and $V = 100$ and sliding the observation window by 1. We evaluate the performance of the WiFiTuned's SHARPA model in semi-controlled and in-the-wild settings.

6.1.1 Head Gesticulation Recognition in Semi-controlled Setup. The Tx and Rx are placed $2m$ apart in a semi-controlled setup, and the target user (T) sits in the line of sight (LOS) as shown in Fig. 7. We compared the performance of the proposed SHARPA model with other state-of-art models such as *SHARP* [35], *Bi-LSTMA* [48], *Constructive model*, *XGBoost (XGB)*, *Contrastive model*,

Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF) and *Multiclass Logistic Regression (LR).* We allocate 60% Doppler traces for training to train the recognition models, 20% for validation and 20% for testing. The performance of the recognition model is shown in Table 3. The SHARPA model outperforms other recognition models with an average improvement of 40.80%/48.32% in the accuracy/F1-score. Moreover, the SHARPA model outperforms the Bi-LSTMA and SHARP models with a 4.68%/4.42% improvement in the accuracy/F1-score. The model recognizes each head gesticulation with an average F1-score of 97.71% (min 89.13%). Next, we train the model 10 fold cross-validation (each time, utilize nine fold as training and one fold as testing). The mean accuracy of SHARPA is 98.3% (better than others). The SHARPA model outperforms other models as shown in Table 3. We further validate the SHARPA model using one participant out cross-validation, trained with 21 participants, and tested on 1. It achieves an average F1-score of 92.36% across all participants. Thus, we conclude the SHARPA model is robust across participants. We further check using one location out cross-validation (each time CSI data of one location is removed from the training set and used as validation) and obtain an average F1-score of 91%. Hence, the model is robust across four locations. Moreover, we train and test the existing SHARPA model using Doppler vector traces computed from the sanitized CSI data using only the phase sanitization procedure. In one location out, the SHARP model could achieve only 82% F1-score. However, after applying additional denoising procedures such as the Hample filter, 1-D wavelet transform filter, and Savitzky-Golay filter, the model achieves a 9.76% improvement in the F1-score. This concludes that only a phase sanitization procedure is not sufficient to clean the CSI data.

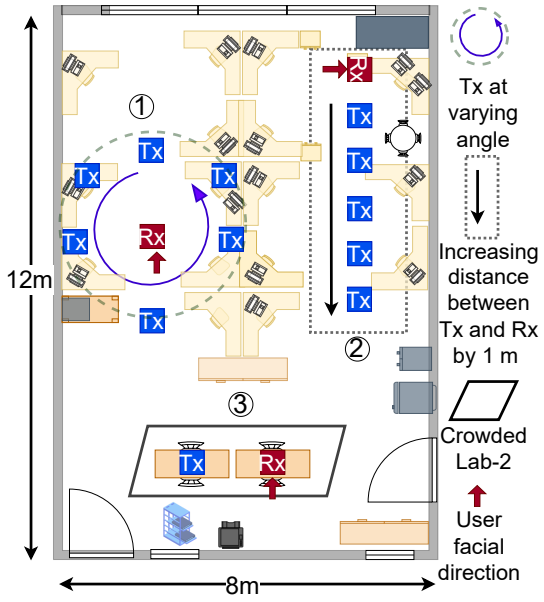
Key Takeaway 1:

Doppler vector traces computed from denoised CSI data provide robust independent features for activity recognition across different indoor environments. The SHARPA model outperforms other classification models and achieves an average F1-score of 92% across all participants and locations.

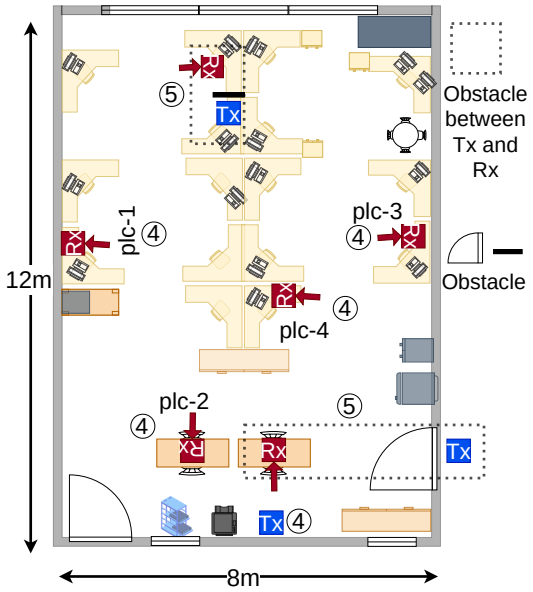
Table 4. Performance of recognition models in in-the-wild settings in F1-score

Model	Changing angle	Increasing distance	Crowded environment	Tx and Rx in different location	Obstacle between Tx and Rx	Tx wall mounted	Interferer near to the target
SHARPA with ATTENTION layer	86.1%	87.44%	84.40%	87.4%	79.11%	80.20%	74.00%
SHARP without ATTENTION layer	80.00%	76.14%	72.00%	67%	71.00%	73.18%	45.00%

6.1.2 Head Gesticulations Recognition in-the-Wild Setup. Compared to the semi-controlled settings, in-the-wild settings are challenging, and the datasets collected in these settings contain different characteristics, nuances, or variations. Consequently, the model parameters need fine-tuning to adapt to nuances or variations inherent in the in-the-wild dataset. We train the state-of-the-art SHARP model with the semi-controlled dataset, and the model achieves an F1-score of 94%. However, when retrained by taking 5% in-the-wild data, it did not adapt well and achieved an F1-score of 69.18%. Hence, the SHARP model is not suitable to deploy in in-the-wild settings. After incorporating the ATTENTION layer (details in §. 4.3), we retrained the SHARPA model by taking 5% of in-the-wild data of each setting. After retraining, the SHARPA model recognizes the head gesticulations with an average F1-score of 86%. This exhibits its adaptability to diverse environments, even with limited training data and achieves a 24.31% improvement in F1-score. We further evaluate the performance of the SHARPA model in each of the in-the-wild settings and compare it with the vanilla SHARP model (without the ATTENTION layer), as shown in Table 4.



(a) Lab-2: **1)** The Tx is positioned at various angles at a distance of $2m$ from the Rx. **2)** The Tx is placed at increasing distances ($1m$ to $5m$) from the Rx by $1m$. **3)** The Tx is placed at a distance of $2m$ in crowded scenarios where more than 10 people are moving around. *The position of the Rx is fixed, as shown in the figure.*



(b) Lab-2: **4**) When the Tx is wall-mounted, the Rx is positioned at various distances ranging from $2m$ to $5m$. **5** Obstacles, such as the nylon cloth sheets, a glass sheet, a wooden panel, and a wooden door, are placed between the Tx and Rx.

Fig. 8. In-the-wild experiment setup in Lab-2.

1. Impact of changing the angle between Tx and Rx: We keep the distance between Tx and Rx as $2m$ but rotate the Tx at radius of $2m$ from the Rx such that the angle between them varies such as 0°, 45°, 90°, 135°, 180° and 270° as shown in Fig. 8a. At 0°, T sits in LOS. The SHARPA models achieve an F1-score of 96.00% at 0°. However, T does not sit in the LOS at other angles, and the SHARPA model achieves an average F1-score of 84.2%. Such an observation highlights the importance of the participant's position in terms of the angle between Tx and Rx. Across all angles, SHARPA achieves an average F1-score of 86.1% as shown in Fig 10a. In comparison, the vanilla SHARP model obtained an average F1-score of 80%. Thus, the SHARPA model demonstrates a notable 7.62% improvement in F1-score compared with SHARP.

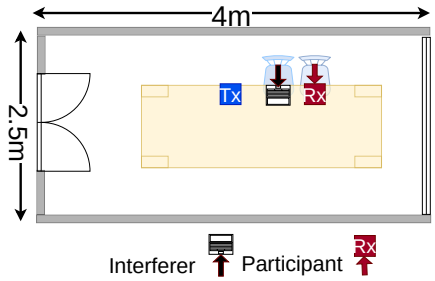


Fig. 9. In-the-wild experiment setup in the meeting room. The interferer (I) sits near the target user (T). The distance between Tx and Rx is set to $2m$.

2. Impact of increasing distance between Tx and Rx: We initially set the distance between Tx and Rx to $1m$. Subsequently, we increase the distance to $5m$ with an increment of $1m$ at a time, as shown in Fig 8a. The Tx and Rx are within LOS of each other. The SHARPA model is trained using the semi-controlled dataset, where the distance between the Tx and Rx is fixed at $2m$. Hence, even in the wild, it achieves 100% F1-score where the distance between Tx and Rx is set to $2m$. With an increasing distance between the Tx and the Rx, the WiFi CSI signal quality (amplitude) decays exponentially.

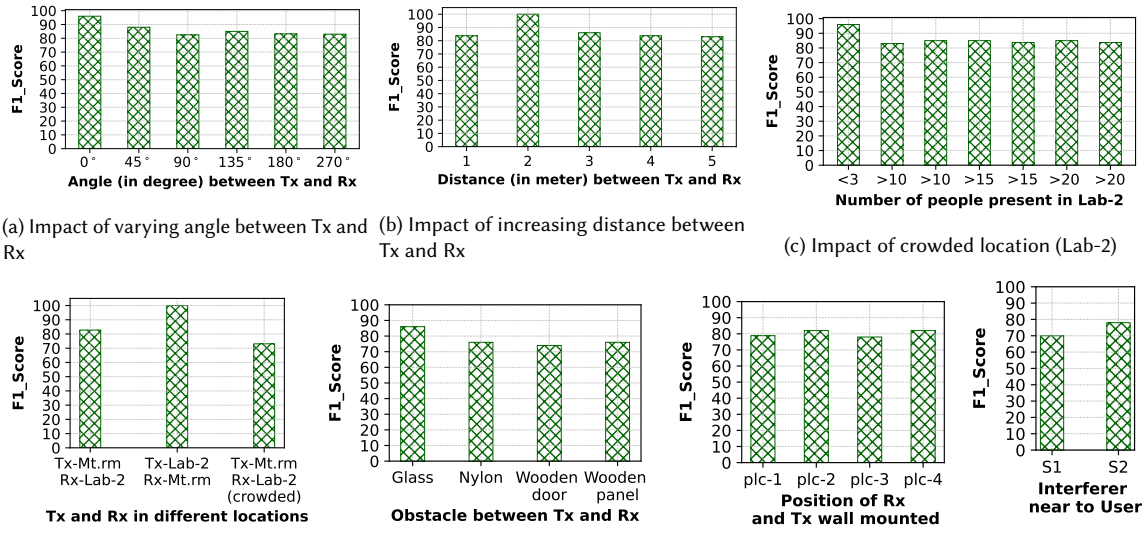
Further, it increases the likelihood of the presence of obstacles and reflection surfaces. Thankfully, we compute the Doppler phase vector from phase and do not rely on amplitude, thus our SHARPA model demonstrates adaptability to increasing distances, achieving an average F1-score of 87.44% as shown in Fig. 10b. In contrast, the vanilla SHARP model obtained an F1-score of 76.14%, which shows the importance of using the ATTENTION layer to discard unstable features.

3. Impact of crowded environment: We conducted experiments when $10-20$ persons moved around, as shown in Fig. 8a. The distance between Tx and Rx is fixed at $2m$, and T sits in LOS. The SHARPA model achieves an average F1-score of 84.40% when the number of people present is more than 3, as shown in Fig 10c. However, when a few people (less than 3) were moving around in the same settings, the SHARPA model achieved an F1-score of 96.12%. On this setup, even the vanilla SHARP model performs well and achieves an F1-score of 90%. However, the SHARP model fails to detect correct head movement in a crowded environment, achieving an F1-score of only 72%. Thus, a crowded environment impacts the CSI data due to the presence of more dynamic components in the environment, and the SHARPA model adapts well and performs better than the SHARP model.

4. Impact of Tx and Rx in different locations: We place Tx in one location (Lab-2/meeting room) and Rx in another location (meeting room/Lab-2). Lab-2 and the meeting room are adjacent to each other. The model achieves an average F1-score of 85.28%. When Rx is placed in the meeting room, the model achieves an F1-score of 99%. However, when Rx is placed at Lab-2, the model achieves an F1 score of 77.92%. Thus, the placement of Rx in crowded and cluttered locations impacts the performance of the SHARPA model when the Tx is in another location. Additionally, the vanilla SHARP model archives an average F1-score of 67%. Thus, the SHARPA model achieves a significant average improvement of 30.40% in the F1-score. This shows that the SHARPA model with the ATTENTION layer adapts well even when Rx is in another location.

5. Impact of obstacle between Tx and Rx: We place obstacles such as a wooden panel (2 inches wide), nylon cloth sheets (4 sheets), a glass sheet (0.25 inches wide), and a wooden door (1 inch wide) between Tx and Rx, as shown in Fig 8b. This experimental setup is challenging, as due to obstacles, the WiFi signal faces huge attenuation. Note that here, the amplitude of the WiFi signal will be attenuated significantly. Thus, any model that detects head gesticulation using amplitude would fail miserably here. In this case, the SHARPA model achieves an F1-score of 76%, as shown in Fig 10e. In the case of the nylon cloth sheets, the model achieves an F1-score of 86.00%. However, in the case of a glass sheet, a wooden panel, and a wooden door, the SHARPA model achieves an average F1-score of 75.33%. The vanilla SHARP without ATTENTION layer struggles to adapt and achieve an average F1-score of 71%. Thus, the material type and width significantly impact the Channel State Information (CSI), thereby impacting the performance of the model. SHARPA outperforms the vanilla SHARP model with an average 6.06% improvement in F1-score. This shows that the ATTENTION layer improves the performance of the recognition model in the presence of obstacles.

6. Impact of wall-mounted Tx: To represent a typical WiFi deployment, where the access point is mounted on the wall, we mount the Tx on the wall above from Rx. Fig. 8b shows where the participant sits at 4 different places i.e., $plc-1, plc-2, plc-3$ and $plc-4$ in Lab-2. The distance and angle between Tx and Rx vary for each place. At $plc-2$ and $plc-3$, the SHARPA model achieves an average F1-score of 82.00%. In both positions, Rx is kept close to Tx, and T sits in LOS. However, at $plc-1$ and $plc-4$, the SHARPA model achieves an average F1-score of 78.4% as shown in Fig. 10f. Rx is placed at a distance greater than $5m$ in both positions, and T is not in LOS. Thus, the distance, angle, and



(a) Impact of varying angle between Tx and Rx

(b) Impact of increasing distance between Tx and Rx

(c) Impact of crowded location (Lab-2)

(d) Impact of Tx and Rx in different locations.

(e) Impact of obstacle type between Tx and Rx.

(f) Impact of different position of Rx w.r.t. wall mounted Tx

(g) Impact of interferer near to target user.

Fig. 10. Performance of *SHARPA* in in-the-wild settings to recognize different head gesticulations

LOS position between the Tx and the Rx impact the robustness of the SHARPA model. However, the SHARPA model outperforms the vanilla SHARP model with an average improvement of 9.59% in the F1-score.

7. Impact of interferer near the target user: To represent a use case where two users might be attending online lectures sitting side by side, we let another user that we call as interferer I sit close to T, and Tx and Rx are placed in the meeting room, as shown in Fig. 9. The distance between Tx and Rx is set to $2m$. The data is collected in two scenarios, $S1$ and $S2$. In $S1$, I performs similar head gesticulations as T. Meanwhile, in $S2$, I performs different head gesticulations from T. In $S1$, the model achieves an F1-score of 70%. The patterns of T's head gesticulations got mixed with I. Therefore, the denoising and learning procedures could not extract relevant patterns that are only related to T. In $S2$, the model performs better and achieves an F1-score of 78%. The SHARPA model outperforms the vanilla SHARP model, achieving a significant average improvement of 64.44% in the F1-score and recognising head gesticulations with 74% F1-score. Such observations highlight that activity recognition is challenging in multi-user environments and needs separate attention to develop specific sensing models that work in such cases [22, 24].

Key Takeaway 2:

The proposed SHARPA model with ATTENTION layers adapts well to various challenging in-the-wild settings that depict real-world scenarios. It makes the WiFiTuned applicable to real-world scenarios in monitoring the engagement of the participants through WiFi signal.

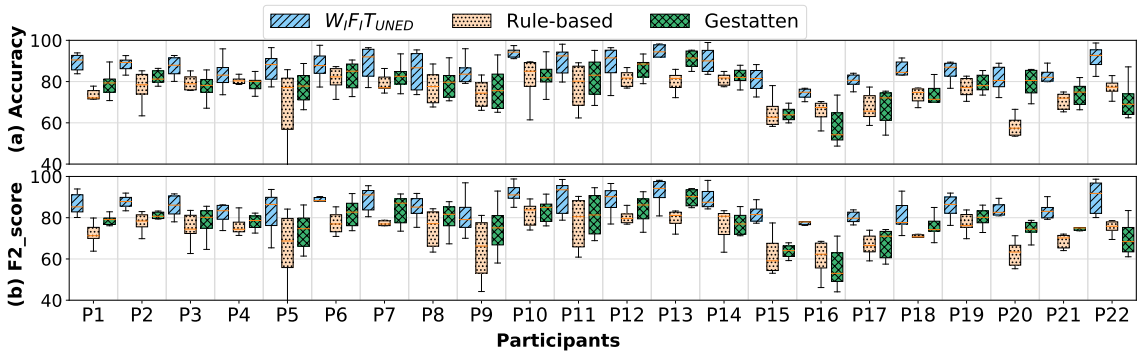


Fig. 11. Participant-wise dist. (distribution) of (a) Accuracy and (b) F2_score at segment level.

6.2 RQ2: Participant-wise Engagement Monitoring

As shown in Fig. 11, WiFiTuned correctly classifies each segment and achieves an average accuracy and F2_score of 86.82% and 85.70% respectively. Gestatten classifies each segment using gaze direction (using the frontal video), which is similar to *looking forward*. However, there can be other gestures of engagement. For example, if a participant listens with full attention but with eyes closed, Gestatten would mark him disengaged. Thus, it suffers from the limitations of video-based inferences. Gestatten achieves an average accuracy of 77.65% and F2_score of 77.09%. The Rule-based baseline uses static rules that cannot correctly classify the segments, as the same physiological and behavioural rules might not hold true for different participants. Thus, the Rule-based baseline lacks robustness and achieves an average accuracy of 74.13% and F2_score of 72.13%. Overall, WiFiTuned shows an average improvement of 11.11%/16.39% in accuracy and 11.39%/18.39% in F2_score over Gestatten/Rule-based. It implies that WiFiTuned overcomes the limitations of the video-based and static rules-based inferences.

Next, we analyze how accurately WiFiTuned computes the engagement score of participants for the entire online meeting. We have a total of 132 engagement scores from 132 meetings ($22 * 6$). We compute the difference ($d_i \in \{d_1, d_2, d_3, ...d_{132}\}$) between generated and ground truth engagement score for each meeting. For this, we perform hypothesis testing with one-sample t-test [19]: ***Null hypothesis (H0):*** The difference d_i between the computed engagement score and ground truth engagement score is significant. ***Alternative hypothesis (H1):*** The difference d_i between the computed engagement score and ground truth engagement score is not significant. We obtain a p-value of 0.02 (WiFiTuned), 0.33 (Rule-based), and 0.35 (Gestatten). For WiFiTuned, p-values **<0.05**, hence we reject *H0* and accept *H1* hypothesis. For Rule-based and Gestatten, p-value> 0.05, hence can not reject the null hypothesis (*H0*). It implies that WiFiTuned matches the ground truth.

Fig. 12 shows the WiFiTuned's capability of correctly distinguishing participants with low, medium, and high engagement scores. We observe that WiFiTuned generated scores very well match with the ground truth, with an average error of 5.86 (range: 2.0 − 16.49). The average error for Gestatten and the Rule-based model are 11.39 (range: 1.06 − 73.65) and 13.07 (range: 1.88 − 71.00), respectively. From Fig. 12, we make the following inferences:***(1) Identification of engaged participants:*** WiFiTuned correctly identifies highly engaged participants (P1-P14) with an average error of 5.92. The generated engagement score ranges from 73.52 to 100. The participants were highly engaged, and the most frequent task was looking at the screen. Thus, Gestatten/Rule-based also works well to identify engaged participants with an average error of 10.91/12.06. However, both provide a low engagement score for participants

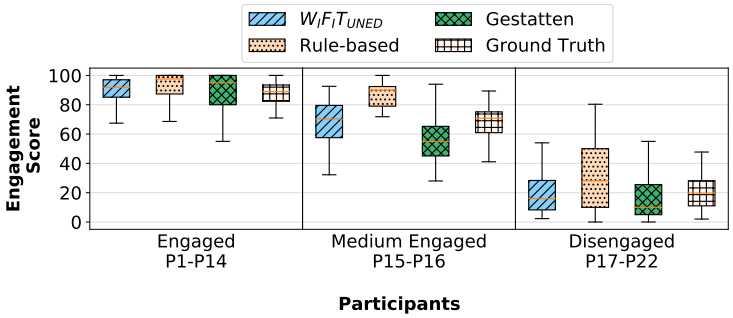


Fig. 12. Dist. of the engagement scores at meeting level.

who were taking notes. ***(2) Identification of participants with medium engagement level (multitaskers):*** For participants P15 and P16, the score computed by WiFiTuned varied from 55.88 to 86.84 and matches the ground truth with an average error of 6.20. The medium-ranged engagement score resulted from the fact that the users were partially attentive while performing periodic parallel tasks, such as frequently taking notes, texting, and using mobile. Gestatten and Rule-based fail to capture multitasking and hence incur an average error of 16.33 and 17.82, respectively. Our hierarchical approach of grouping segments at the intent level allows for determining positive or negative multitasking for momentary or longer distractions. This allows WiFiTuned to identify different types of engagement levels across participants accurately. ***(3) Identification of disengaged participants:*** WiFiTuned correctly identifies all the disengaged (P17-P22) participants and generates their engagement scores in the range of 2.3 and 40.34 and incurs an average of 5.71. Gestatten and Rule-based model computes the engagement score with an average error of 10.33 and 15.76.

Key Takeaway 3:

WiFiTuned accurately determines positive and negative multitasking for momentary and longer distractions. Moreover, WiFiTuned adapts to the subjective differences in the behaviour of participants attending online meetings. Thus, WiFiTuned can provide reliable monitoring of the engagement level of the participants.

6.3 RQ3: Meeting Content-wise & Location-wise Engagement Monitoring

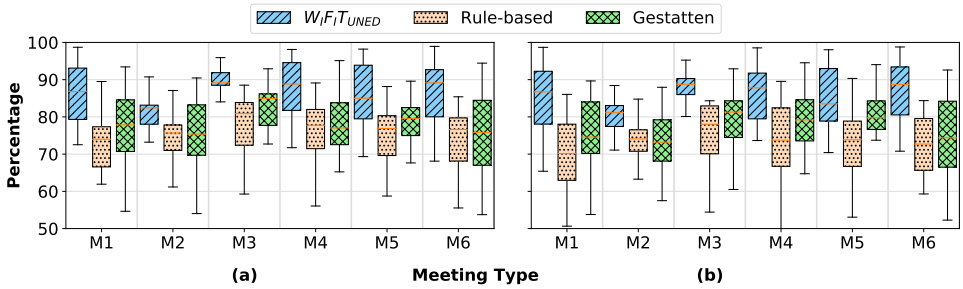


Fig. 13. Content-wise dist. of (a) Accuracy, (b) F2_score at segment level.

6.3.1 ***Content-wise Engagement Analysis.*** Fig. 13 shows the distribution of content-wise accuracy and F2_score (content type details in §. 5.1). For meeting M1 and M3-M6, WiFiTuned performs better than baseline models with an average improvement of 11.31%/17.93% in accuracy and 11.59%/19.42% in F2_score over Gestatten/Rule-based. For M1, The engaged participants (as per ground truth annotation) nod/shake their heads as feedback and close their eyes while thinking about the questions. For M3 and M5, the engaged participants look at the screen and understand the concepts. The participants also follow the speaker's command. For M4, the participants listen actively and follow the content. The observation of M6 is similar to M5. The duration of M2 was 45 min, the attention span was not throughout the online meeting. The participants look around for a short duration, take notes, and nod/shake their heads while listening and giving feedback. Further, the participants change their positions frequently, such as lying backwards on the chair and leaning forward on the desk as they sigh. WiFiTuned could not adapt well here as participants' sitting postures change frequently. It achieves an accuracy/F2_score of 80.87%/79.71%. WiFiTuned needs to adapt to such changing postures, still outperforms the baselines by 8.4%/8.2% in accuracy and 9.67%/9.09% in F2_score over Gestatten/Rule-based. WiFiTuned monitors the engagement level in the context of the speaker's intent and contextualizes head gesticulations and intent. By not merely counting the frequency of expected or unexpected head gesticulations, our model extracts the alignment of head gesticulation pattern and intent. Thus, it could adapt well to different users. As a result, WiFiTuned provides robust engagement monitoring even across different meeting content.

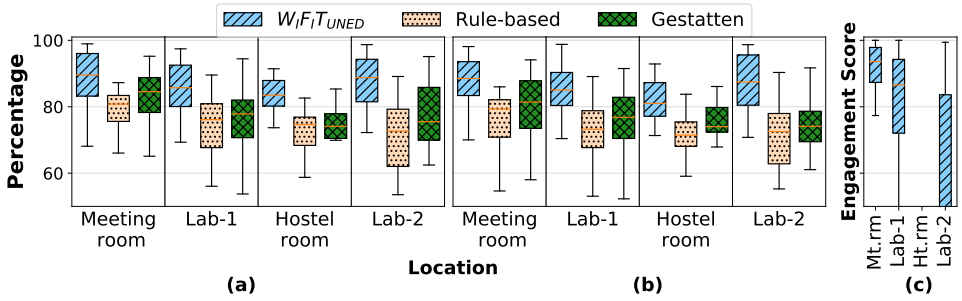


Fig. 14. Location-wise dist. of (a) Accuracy, (b) F2_score at segment level, and (c) Engagement score at meeting level.

6.3.2 ***Location-wise Engagement Analysis.*** Fig. 14(a) shows the distribution of location-wise accuracy and F2_score. All the participants who joined the online meetings from the meeting room were engaged. In Lab-1, all participants were engaged except one (using the smartphone frequently). In both locations, WiFiTuned shows an average improvement of 9.805%/15.33% in accuracy and 9.93%/14.96% in F2_score over Gestatten/Rule-based. In the hostel room, participant P17 joins in a lying position. As his head movements get restricted, WiFiTuned could not compute the engagement score correctly (accuracy 79.64% and F2_score 78.34%). P18 and P19 join the online meeting in the sitting position. WiFiTuned computes the engagement score with an average improvement of 12.08%/15.70% in accuracy and 10.44%/13.56% in F2_score from Gestatten/Rule-based. Participants who joined from Lab-2 got distracted by fellow lab mates, frequently used their mobile, lying on the chair, and were involved in other irrelevant tasks such as reading irrelevant papers. Such a distraction was momentary for some users and sustained for the rest. WiFiTuned ignores momentary disengaged behaviour, considering the fact a participant can not be engaged all the time. However, *Gestatten* does not consider momentary disengagement and generates a low engagement score for all. The rule-based model also could not adapt well to situations where participants were involved in multitasking. WiFiTuned correctly classify the segment with

an average improvement of 13.74%/21.01% in accuracy and 16.08%/23.08% in F2_score compared Gestatten/Rule-based. Fig. 14 (b) shows the distribution of location-wise engagement scores: (1) meeting room (76.60 – 100), (2) Lab-1 (56.60 – 100), (3) hostel room (2.3 – 35.65), and (4) Lab-2 (3.70 – 99.41). The environment/location impacts the engagement behaviour of the participants.

Key Takeaway 4:

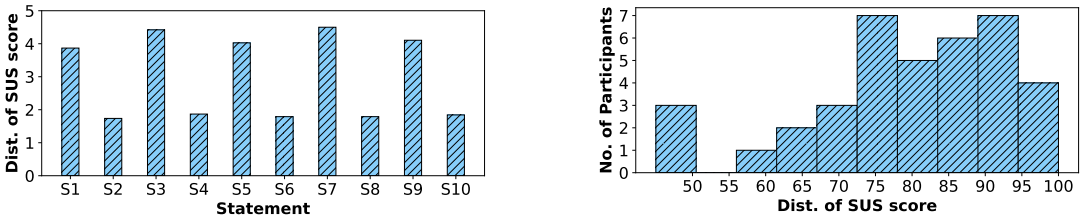
The environment/location and meeting content type and length significantly affect the engagement level of the participants. The engagement level includes various states, such as actively listening, multitasking, momentary distraction, and disengagement. To capture these states, WiFiTuned monitors the engagement level in the context of the speaker's intent. It contextualizes head gesticulations and intent to determine whether a participant is fully engaged, medium engaged or disengaged. Thus, WiFiTuned provides comprehensive and accurate monitoring of the engagement level of the participant throughout the online meeting.

6.4 Running Time

We used a standard laptop with Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz with 8 GB RAM to infer the runtime information. For a 10s segment, the preprocessing and denoising together take 0.5s and extracting spoken utterances takes 1s. The trained head gesticulation model takes 0.018s, and the trained intent recognition model takes 0.0015s. The engagement monitoring module takes 0.0003s to classify the segment. In total, WiFiTuned takes 1.6s for each 10s segment. Similarly, the *Rule-based* model also takes 1.6s. However, *Gestatten* takes 10.05s to generate the engagement score.

Key Takeaway 5:

WiFiTuned takes 1.6*s* to classify 10*s* meeting segments as engaged or disengaged with a standard laptop. This makes the WiFiTuned applicable for real-time monitoring



(a) Statement-wise score distribution.

(b) Participant-wise score distribution.

Fig. 15. Statement-wise and Participant-wise dist of SUS score.

6.5 Usability Study

A total of 43 participants participated in the usability study of WiFiTuned. Out of them, 10 participants of 20 – 30 years of age participated in the study physically by attending online classes using our setup. These participants are

independent of those who were involved in data collection. Each participant joins the online meeting with a setup as shown in Fig. 6 with WiFiTuned for 5 minutes. After the meetings, we provided each participant with their engagement score for self-annotation (engaged or disengaged). Additionally, we made a demo video that demonstrates the working of WiFiTuned. Another 33 participants, $20-30$ years of age, provided feedback by watching the demo video. Thus, a total of 43 participants provided feedback on the usability of WiFiTuned. We have used the standard questionnaire from the system usability scale (SUS) [3], where the participants provide their feedback on the scale of 1 (*strongly disagree*) to 5 (*strongly agree*) for 10 different usability statements. Notably, strong agreement towards odd statements and strong disagreement towards even statements indicate the high usability of WiFiTuned. We calculate the scaled average score using the formula shown in [3] for each participant.

Fig. 15 shows the distribution of the SUS score obtained as participant feedback from both the real deployment and demo video (Cronchbach alpha=0.73). As shown in Fig. 15a, the negative statements have received the lower scores, and the positive statements have received the higher scores. We get an average SUS score of 79.73%, indicating that participants, on average, felt WiFiTuned as a usable system. Fig. 15b shows that the majority of the participants have given SUS scores of more than 70.

Key Takeaway 6:

WiFiTuned attains an average usability score of 79.73% on the system usability scale (SUS), which indicates the overall adaptability of the platform for real-world usage.

7 DISCUSSION AND LIMITATION

Through the extensive evaluation of WiFiTuned, we found it to be quite effective and accurate under different environmental setups. Further, the evaluation resulted in some significant observations and limitations. We discuss them and mention the scope for future works addressing them.

Positional changes: WiFiTuned's performance has been analyzed for participants attending the online meeting in static positions (sitting, lying, and leaning). Their positions were fixed. Reducing body movements can enhance engagement and promote focused listening. However, in an online scenario, the users can move freely in their personal space. For example, in the absence of a formal meeting setup, a user might walk around the room while attending the meeting. In such a scenario, the head gesticulations and the signatures associated with the individual gesticulation type might vary widely. Future works will aim at handling such robust scenarios by utilizing additional modalities like indoor location tracking, along with the individual's head gesticulations.

Ambience and Engagement: The evaluation of WiFiTuned has shown that the environmental ambience plays a crucial role in determining the engagement level of participants. The participants are more prone to distractions in crowded indoor environments, and the presence of multiple users can impact the head gesticulation recognizer's performance. Interference from external sources can affect the CSI and thus impact performance. In this paper, we have retrained the SHARPA model to adapt to in-the-wild settings using only 5% of the in-the-wild CSI dataset. To make WiFiTuned more reliable and efficient in real-world scenarios, future research will focus on utilizing only a few labelled CSI samples to enable the SHARPA model to quickly adapt to new environment settings utilizing few-shot or one-shot embedded

learning [53, 56].

Human behaviour: Human behaviour is diverse in nature. Hence, head gesticulations can largely vary for people having cultural and geographical differences. Moreover, the participants for the evaluation of WiFiTuned are mostly young people belonging to the age group of 20 – 30 years. Future directions will aim at involving people from other age groups to test the performance of WiFiTuned. Moreover, users with clinical challenges like Essential tremors, Parkinson's disease, and others might experience involuntary and rapid head tremors. In such cases, WiFiTuned might generate false positives and incorrectly identify them as disengaged.

8 CONCLUSION

In this paper, we propose a multimodal engagement monitoring system WiFiTuned. WiFiTuned tracks and recognizes different types of user's head gesticulations through WiFi CSI, providing a non-intrusive, passive, and privacy-preserving sensing modality. Further, the system utilizes the intent of the speaker's speech and correlates it with the head gesticulations of the listeners as they attend these online meetings. We then use such a correlation to infer the listener's engagement in online meetings. We evaluate WiFiTuned with 22 participants in 4 different locations with 6 different meeting types in varied setups. Our extensive evaluation under varied setups reveals the significant efficacy of WiFiTuned and its promising applicability in monitoring engagement in online meetings or providing attendees with insights into their engagement patterns. Our in-the-wild evaluation shows that WiFiTuned is adaptable and scalable. The human studies show the system is user-friendly, scoring an overall system usability score of 79.73%.

REFERENCES

[1] Arqam M Al-Nuimi and Ghassan J Mohammed. 2021. Face Direction Estimation based on Mediapipe Landmarks. In *ICCITM 2021*. IEEE, 185–190.

[2] Mohammud J Bocus, Kevin Chetty, and Robert J Piechocki. 2021. UWB and WiFi systems as passive opportunistic activity sensing radars. In *2021 IEEE Radar Conference (RadarConf21)*. IEEE, 1–6.

[3] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[4] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. 2007. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE transactions on audio, speech, and language processing* 15, 3 (2007), 1075–1086.

[5] Hancheng Cao, Chia-Jung Lee, Shamsi Iqbal, Mary Czerwinski, Priscilla NY Wong, Sean Rintel, Brent Hecht, Jaime Teevan, and Longqi Yang. 2021. Large scale analysis of multitasking behavior during remote meetings. In *CHI 2021*. 1–13.

[6] Jun-Ho Choi, Marios Constantinides, Sagar Joglekar, and Daniele Quercia. 2021. KAIROS: Talking heads and moving bodies for successful meetings. In *Proc. HOTMOBILE 2021*. 30–36.

[7] Fabiola Colone, Francesca Filippini, Marco Di Seglio, Paul V Brennan, Rui Du, and Tony Xiao Han. 2023. Reference-free amplitude-based WiFi passive sensing. *IEEE Trans. Aerospace Electron. Systems* 59, 5 (2023), 6432–6451.

[8] Marco Cominelli, Francesco Gringoli, and Francesco Restuccia. 2023. Exposing the CSI: A Systematic Investigation of CSI-based Wi-Fi Sensing Capabilities and Limitations. In *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 81–90.

[9] Marco Cominelli, Francesco Gringoli, and Francesco Restuccia. 2023. Exposing the csi: A systematic investigation of csi-based wi-fi sensing capabilities and limitations. In *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 81–90.

[10] Snigdha Das, Sandip Chakraborty, and Bivas Mitra. 2021. Quantifying Students' Involvement during Virtual Classrooms: A Meeting Wrapper for the Teachers. In *India HCI 2021*. 133–139.

[11] Berardina De Carolis, Francesca D'Errico, Nicola Macchiarulo, and Giuseppe Palestra. 2019. "Engaged Faces": Measuring and Monitoring Student Engagement from Face and Gaze Behavior. In *IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume* (Thessaloniki, Greece) *(WI '19 Companion)*. Association for Computing Machinery, New York, NY, USA, 80–85. https://doi.org/10.1145/3358695.3361748

[12] Triparna de Vreede, Stephanie A Andel, Gert-Jan de Vreede, Paul Spector, Vivek Singh, and Balaji Padmanabhan. 2019. What is engagement and how do we measure it? Toward a domain independent definition and scale. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[13] M Ali Akber Dewan, Fuhua Lin, Dunwei Wen, Mahbub Murshed, and Zia Uddin. 2018. A deep learning approach to detecting engagement of online learners. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 1895–1902.

[14] Jianyang Ding, Yong Wang, Hongyan Si, Shang Gao, and Jiwei Xing. 2022. Three-dimensional indoor localization and tracking for mobile target based on wifi sensing. *IEEE Internet of Things Journal* 9, 21 (2022), 21687–21701.

[15] Yi Fang, Wei Liu, and Sun Zhang. 2023. Wi-Senser: Contactless Head Movement Detection during Sleep Utilizing WiFi Signals. *Applied Sciences* 13, 13 (2023), 7572.

[16] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. 2020. n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–26.

[17] Ruiyang Gao, Wenwei Li, Yaxiong Xie, Enze Yi, Leye Wang, Dan Wu, and Daqing Zhang. 2022. Towards Robust Gesture Recognition by Characterizing the Sensing Quality of WiFi Signals. *Proc. Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–26.

[18] Yang Gao, Yincheng Jin, Seokmin Choi, Jiyang Li, Junjie Pan, Lin Shu, Chi Zhou, and Zhanpeng Jin. 2021. SonicFace: Tracking Facial Expressions Using a Commodity Microphone Array. *Proceedings of the ACM on IMWUT* 5, 4 (2021), 1–33.

[19] Banda Gerald. 2018. A brief review of independent, dependent and one sample t-test. *International journal of applied mathematics and theoretical physics* 4, 2 (2018), 50–54.

[20] Reza Hadi Mogavi, Yankun Zhao, Ehsan Ul Haq, Pan Hui, and Xiaojuan Ma. 2021. Student barriers to active learning in Synchronous online classes: Characterization, reflections, and suggestions. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*. 101–115.

[21] Luke Haliburton, Svenja Yvonne Schött, Linda Hirsch, Robin Welsch, and Albrecht Schmidt. 2023. Feeling the Temperature of the Room: Unobtrusive Thermal Display of Engagement during Group Communication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–21.

[22] Khandaker Foysal Haque, Milin Zhang, and Francesco Restuccia. 2023. Simwisense: Simultaneous multi-subject activity classification through wi-fi signals. In *2023 IEEE 24th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 46–55.

[23] Steven M. Hernandez and Eyuphan Bulut. 2020. Lightweight and Standalone IoT Based WiFi Sensing for Active Repositioning and Mobility. In *WoWMoM 2020*. Cork, Ireland.

[24] Jingzhi Hu, Tianyue Zheng, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. MUSE-Fi: Contactless muti-person sensing exploiting near-field Wi-Fi channel variation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[25] Amelia R Hunt and Alan Kingstone. 2003. Covert and overt voluntary attention: linked or independent? *Cognitive Brain Research* 18, 1 (2003), 102–105.

[26] Stephen Hutt, Caitlin Mills, Nigel Bosch, Kristina Krasich, James Brockmole, and Sidney D'mello. 2017. " Out of the Fr-Eye-ing Pan" Towards Gaze-Based Models of Attention during Learning with Technology in the Classroom. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. 94–103.

[27] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. 2018. Engagemon: Multi-modal engagement sensing for mobile games. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 1 (2018), 1–27.

[28] Hui Jiang, Chong-Wah Ngo, and Hung-Khoon Tan. 2006. Gestalt-based feature similarity measure in trademark database. *Pattern recognition* 39, 5 (2006), 988–1001.

[29] Pragma Kar, Samiran Chattopadhyay, and Sandip Chakraborty. 2020. Gestatten: Estimation of User's Attention in Mobile MOOCs From Eye Gaze and Gaze Gesture Tracking. *Proc. ACM Hum.-Comput. Interact.* 4, EICS (2020), 1–32.

[30] Pragma Kar, Shyamvanshikumar Singh, Avijit Mandal, Samiran Chattopadhyay, and Sandip Chakraborty. 2023. ExpresSense: Exploring a Standalone Smartphone to Sense Engagement of Users from Facial Expressions Using Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 265, 18 pages. https://doi.org/10.1145/3544548.3581235

[31] Anastasia Kuzminykh and Sean Rintel. 2020. Classification of functional attention in video meetings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[32] Andrew Lepp, Jacob E Barkley, Aryn C Karpinski, and Shweta Singh. 2019. College students' multitasking behavior in online versus face-to-face courses. *Sage Open* 9, 1 (2019), 2158244018824505.

[33] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An Adaptable System for Instructors to Grasp Student Learning Status in Synchronous Online Classes. In *CHI Conference on Human Factors in Computing Systems*. 1–25.

[34] Yongsen Ma, Gang Zhou, and Shuangquan Wang. 2019. WiFi sensing with channel state information: A survey. *ACM Computing Surveys (CSUR)* 52, 3 (2019), 1–36.

[35] Francesca Meneghello, Domenico Garlisi, Nicolò Dal Fabbro, Ilenia Tinnirello, and Michele Rossi. 2022. Sharp: Environment and person independent activity recognition with commodity ieee 802.11 access points. *IEEE Transactions on Mobile Computing* (2022).

[36] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D'Mello. 2016. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing* 8, 1 (2016), 15–28.

[37] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Köhler. 2018. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In *ICMI 2018*. 191–199.

[38] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. KLUE: Korean Language Understanding Evaluation. arXiv:2105.09680 [cs.CL]

[39] Ronald K. Pearson, Yrjö Neuvo, Jaakko Astola, and Moncef Gabbouj. 2015. The class of generalized hampel filters. In *2015 23rd European Signal Processing Conference (EUSIPCO)*. 2501–2505.

[40] Rupendra Raavi, Mansour Alqarni, and Patrick CK Hung. 2022. Implementation of Machine Learning for CAPTCHAs Authentication Using Facial Recognition. In *ICDSIS*. IEEE, 1–5.

[41] Yili Ren, Sheng Tan, Linghan Zhang, Zi Wang, Zhi Wang, and Jie Yang. 2020. Liquid level sensing using commodity wifi in a smart home environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.

[42] Arta Rosaen and Lidiman Sinaga. 2012. Speech Function in Feature Stories in Reader's Digest. *Linguistica* 1, 1 (2012), 146423.

[43] Nancy P Rothbard. 2001. Enriching or depleting? The dynamics of engagement in work and family roles. *Administrative science quarterly* 46, 4 (2001), 655–684.

[44] Ognjen Rudovic, Meiru Zhang, Bjorn Schuller, and Rosalind Picard. 2019. Multi-modal active learning from human data: A deep reinforcement learning approach. In *ICMI 2019*. 6–15.

[45] Wilmar B Schaufeli, Arnold B Bakker, and Marisa Salanova. 2006. The measurement of work engagement with a short questionnaire: A cross-national study. *Educational and psychological measurement* 66, 4 (2006), 701–716.

[46] Bidisha Sharma, Maulik Madhavi, and Haizhou Li. 2021. Leveraging acoustic and linguistic embeddings from pretrained speech and language models for intent classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7498–7502.

[47] Tripti Singh, Mohan Mohadikar, Shilpa Gite, Shruti Patil, Biswajeet Pradhan, and Abdullah Alamri. 2021. Attention span prediction using head-pose estimation with deep neural networks. *IEEE Access* 9 (2021), 142632–142643.

[48] Vijay Kumar Singh, Pragma Kar, Ayush Madhan Sohini, Madhav Rangaiah, Sandip Chakraborty, and Mukulika Maity. 2023. WiFiTuned: Monitoring Engagement in Online Participation by Harmonizing WiFi and Audio. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 670–678.

[49] Jingjing Wang, Xianqing Wang, Jishen Peng, Jun Gyu Hwang, and Joon Goo Park. 2021. Indoor Fingerprinting Localization Based on Fine-grained CSI using Principal Component Analysis. In *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*. 322–327.

[50] Yuehua Wang, Anuhya Kotha, Pei-heng Hong, and Meikang Qiu. 2020. Automated student engagement monitoring and evaluation during learning in the wild. In *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)*. IEEE, 270–275.

[51] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[52] Dan Wu, Daqing Zhang, Chenren Xu, Yasha Wang, and Hao Wang. 2016. WiDir: walking direction estimation using wireless signals. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 351–362.

[53] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. Onefi: One-shot recognition for unseen gesture via cots wifi. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 206–219.

[54] Xiang Xiao and Jingtao Wang. 2017. Understanding and detecting divided attention in mobile MOOC learning. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2411–2415.

[55] Matin Yarmand, Jaemarie Solyst, Scott Klemmer, and Nadir Weibel. 2021. "It feels like I am talking into a void": Understanding interaction gaps in synchronous online classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–9.

[56] Guolin Yin, Junqing Zhang, Guanxiong Shen, and Yingying Chen. 2022. Fewsense, towards a scalable and cross-domain wi-fi sensing system using few-shot learning. *IEEE Transactions on Mobile Computing* (2022).

[57] Hao Yu, Ankit Gupta, Will Lee, Ivon Arroyo, Margrit Betke, Danielle Allesio, Tom Murray, John Magee, and Beverly P Woolf. 2021. Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems. In *International Conference on Human-Computer Interaction*. Springer, 219–233.

[58] Nan Yu, Wei Wang, Alex X Liu, and Lingtao Kong. 2018. QGesture: Quantifying gesture distance and direction with WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–23.

[59] Youwei Zeng, Dan Wu, Ruiyang Gao, Tao Gu, and Daqing Zhang. 2018. FullBreathe: Full human respiration detection exploiting complementarity of CSI phase and amplitude of WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–19.

[60] Yong Zhang, Qingqing Liu, Yujie Wang, and Guangwei Yu. 2022. CSI-Based Location-Independent Human Activity Recognition Using Feature Fusion. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–12.

[61] Rongrong Zhu, Liang Shi, Yunpeng Song, and ZhongMin Cai. 2023. Integrating Gaze and Mouse Via Joint Cross-Attention Fusion Net for Students' Activity Recognition in E-learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 145 (sep 2023), 35 pages. https://doi.org/10.1145/3610876

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009