WiFiTuned: Monitoring Engagement in Online Participation by Harmonizing WiFi and Audio

Anonymous Author(s)

ABSTRACT

This paper proposes a multi-modal, non-intrusive and privacy preserving system WiFiTuned for monitoring engagement in online participation i.e., meeting/classes/seminars. It uses two sensing modalities i.e., WiFi CSI and audio for the same. WiFiTuned detects the head movements of participants during online participation through WiFi CSI and detects the speaker's intent through audio. Then it correlates the two to detect engagement. We evaluate WiFiTuned with 22 participants and observe that it can detects the engagement level with an average accuracy of more than 86%.

ACM Reference Format:
Anonymous Author(s). 2018. WiFiTuned: Monitoring Engagement in Online Participation by Harmonizing WiFi and Audio. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Online classes, seminars, and meetings have become increasingly popular despite resuming physical interactions after the Covid-19 pandemic. However, the effectiveness of any interactions, physical or virtual, depends on the active engagement of its participants. Unlike face-to-face interactions, distractions are usual in online meetings due to surrounding objects or persons. Further, multitasking, like taking notes, searching related contents on the Internet, chatting over the mobile, drinking or eating, etc., are common during online meetings [18]. While several existing studies have indicated that multitasking can be effective many a times [4, 17, 22], there is a possibility that a participant might miss a vital content or an important message while involved in multitasking. Imagine a smart virtual meeting platform that can understand its participant's attentiveness and automatically knock them when they miss some critical discussion or selectively save the conversation when the participant is out of focus to replay them later. Such an application is likely to improve the productivity of the meeting or can help its user to utilize the scope of multitasking more positively (like searching for an unknown term on the Internet while not missing any critical discussion).

However, the idea of monitoring user engagement is not new; several prior works [6, 7, 11, 16, 19, 24] have demonstrated efficient approaches to monitor user engagement and even detected instances of multitasking during a virtual meeting. These works primarily use the participant's video feed as the primary modality and utilize techniques like modeling 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 system.

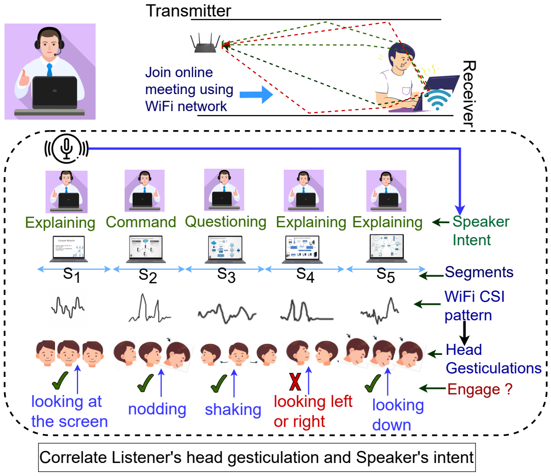


Figure 1: Overall working of WiFiTuned.

This paper considers the fact that nowadays, WiFi is omnipresent in the indoor environment, and most online participants utilize wireless connectivity at the last mile. Notably, in recent years, WiFi sensing using *Channel State Information* (CSI) has emerged as one of the most promising areas for Human Activity Recognition (HAR) [36–38, 40] and Human Gesture Recognition (HGR) [10, 41]. Exploiting the CSI captured over the device being used and correlating it with the speaker's audio in the meeting, we develop a multi-modal interface 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 [3] 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, is likely to show no response when the speaker gives a command. Through an anonymous survey of over 300 participants (having experience

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

with online meetings) from four different countries, we hypothesize that head gesticulations correlate well with the engagement level during online meetings (details in §3). WiFiTuned works on this hypothesis and correlates the listener's head gesticulations 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.

Challenges: However, developing an automated engagement monitoring system has several challenges. (1) Considering engagement is subjective, deciding on an objective evaluation for engagement is not straightforward. (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. (3) Developing a correlation model between the head gestures and the audio intent is challenging as we must provide quantitative ways of correlating them, and there might be multiple different head movements 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 by modeling 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 develop a head-gesticulation recognition model carefully, knowing that WiFi CSI gets impacted by the environment. Accordingly, WiFiTuned attaches an *attention* layer to the neural network model (Bi-directional LSTM) that minimizes the impact of the environment. For correlating head gesticulation and the speaker's intent from the audio, we divide the entire meeting into small fixed-size segments and cluster them in two levels – first, based on the head gesticulation patterns, and then based on the speaker's intent. Next, we determine whether a sub-cluster and associated segments are *engage* or disengage by matching the actual head gesticulations with expected ones for that sub-cluster. 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 a way of monitoring engagement by correlating the speaker's intent and listeners' head gesticulations. We evaluate WiFiTuned with 22 participants in four different locations and for 6 different types of content. We observe that WiFiTuned obtains an average accuracy of more than 86%.

2 RELATED WORK

Prior work has used different types of modalities such as video, audio, and sensor data, to monitor the engagement level. First, we discuss prior work that uses the most common yet single modality i.e., video. **Video based engagement monitoring:** Such models monitor non-verbal aspects, such as facial expressions, eye gaze, head pose, head movements, and hand movements, to monitor the user's attention under different application domains. Das *et al.* [7] and Kar *et al.* [16] have used eye gaze and gaze gestures to obtain student's attention in online classes. Both works investigate whether the participant is following the meeting content or not. While overt attention can be well-tracked by tracking the visual cues like gaze gestures and saccades, covert attention is not reflected by visual behaviors [14]. Moreover, the video-based gazing methods lead to the unfair assessment of attention in the case of intentional blindness [21]. Monkaresi *et al.* [24] utilize facial expressions and heart rates to monitor user's engagement in educational settings. In the same direction, authors in [2, 23, 33, 34] use visible emotional attributes such as bored, anxiety, happiness, confusion, and satisfied to quantify engagement. [5] extract facial expression, head pose, and body pose and employ heuristics rule to monitor engagement in online meetings. Similarly, [32] presents an attention estimation technique that considers the head pose of the subjects. [5] utilize hand movements and body fidgets along with eye gaze, head pose, and facial action units to monitor engagement.

Multi-modal engagement monitoring: Now, we discuss prior work that uses multiple modalities for engagement monitoring. Gao *et al.* [9] estimates emotional, behaviour, and cognitive engagement using two modalities: electrodermal activity (EDA) sensors data and indoor environment data such as temperature, humidity, and sound. The two modalities provide an opportunity to monitor the engagement of participants by exploring physiological, activity, and environmental factors. However, Disalvo *et al.* [8] has revealed that the EDA sensor is not correlated with emotional engagement in classroom settings. Rudovic *et al.* [30] integrates three modalities (audio, video, and EDA sensor) to monitor the engagement of autistic children in real-world child-robot interactions. While some users are more expressive with their voice and some with their expression. Selecting the best modality for individual users is crucial to develop the engagement model. Otsuka *et al.* [25] monitors the gaze direction using non-verbal features such as head pose, eyeball direction, and utterances in multiparty meetings. However, the proposed work does not take into account the audio data, which can provide valuable information.

3 MOTIVATION

We conducted a large-scale anonymous survey to study collective and common head gesticulations 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. 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 perform various type 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/mails"* (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 participant perform certain tasks due to lack of supervision that contributes to disengagement such as "eating food", "using smartphones", "checking mails", "watching videos", "dozing", etc.

3.2 Correlation between Head Nod/Shake and Engagement

We asked the participants if they nod/shake their head 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 Correlation between 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 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%).

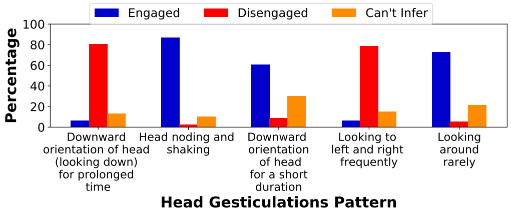


Figure 2: Mapping between different head gesticulation patterns and engagement indicated by the participants.

Further, we asked the participants to provide a rating between 1–5 (1 – *do not reveal engagement*, 5 – *completely reveal engagement*) to head gesticulations pattern. 88.7% participants mostly/always (4 – 5) find head gesticulations highly helpful in indicating the engagement level. Fig. 2 presents a summary of how different types of head gesticulations pattern 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. **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, head gesticulations based system monitors engagement without depending on opening the camera. We utilize these takeaways for designing our system.

4 WIFITUNED DESIGN

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.

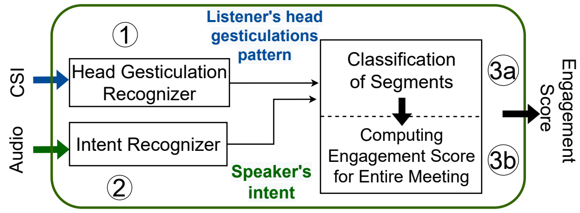

Listener's head gesticulations pattern
1
CSI
Head Gesticulation Recognizer
Classification of Segments
3a
Audio
Intent Recognizer
Computing Engagement Score for Entire Meeting
3b
Speaker's intent
2
Engagement Score

Figure 3: Design of WiFiTuned: Input: WiFi CSI & Audio. Output: Engagement score.

4.1 Recognizing Head Gesticulation from CSI

WiFi CSI (Channel State Information) captures perturbations in WiFi signal caused by obstacles in the environment (such as the human body) in terms of various channel impairments such as reflection, absorption, diffraction, scattering, etc. The CSI ($\mathbb{H}$) is represented as $R = T\mathbb{H}+\eta$, where T and R is transmitted and received signal vector respectively and η is additive white Gaussian noise vector. $\mathbb{H}$ consists two components, real (h^i_{re}) and imaginary (h^i_{img}) for each subcarrier (N^i). The collected CSI data is represented by $R \times N \times A$ matrix, where N is the total number of data subcarriers, R is the total number of samples, and A is the total number of receiving antennas.

4.1.1 Preprocessing, Feature Extraction & Ground truth Labelling. We compute the *amplitude* ($\mathcal{A}$) as a feature of each subcarrier as $\mathcal{A} = \sqrt{(h^i_{img})^2 + (h^i_{re})^2}$. Notably, CSI is impacted by head movements, surrounding objects and people moving around, multi-path effects, hardware/software errors, and processing errors. Thus, we use the denoising procedure explained in [35]. We first apply a Hample filter [27] to remove the high-frequency noise/anomalies then a 1-D Wavelet transform (DWT) filter [39] to remove noise caused by surrounding objects and people moving around. Finally, we use Savitzky_Golay smoothing filter [20] to preserve the fluctuation induced by head gesticulations. We use video data for ground-truth labeling of CSI data with the corresponding head gesticulation labels listed in Table 1. We determine head gesticulations using MediaPipe [1] and PnP algorithm [28]. 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, PnP algorithm uses these landmarks to detect the head pose in 3D space. After that, six 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 sec. For example, nodding involves continuous movement of the head up and down vertically, comprising looking up (down), forward, and down (up) orientations. Similarly, head shaking comprises looking left (right), forward, and looking right (left). Finally, the recognized head gesticulations are used as ground truth and the timestamp is used to synchronize with CSI data. We use the labeled CSI data to train the head gesticulation recognition model.

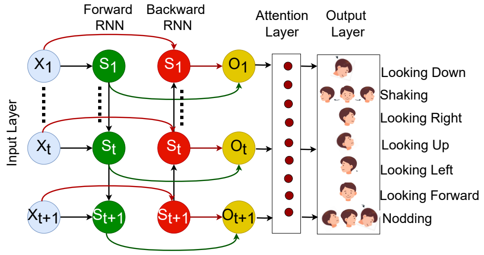


Figure 4: Bi-LSTMA architecture.

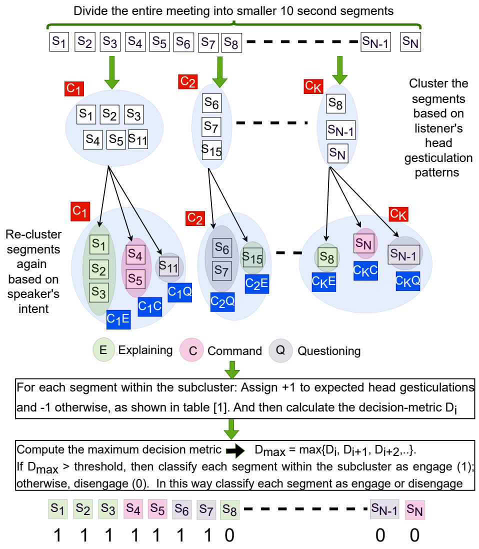
4.1.2 Head Gesticulation Recognizer. A single head gesticulation consists of multiple head orientations that are connected in time. Hence, a recognition model needs not only past head orientations but also future orientations. A bi-directional Long Short-Term Memory model with an attention layer (Bi-LSTMA), can encode temporal information, is a strong option for automatic feature learning. As shown in Fig. 4, both forward and backward processes of feature learning are included in Bi-LSTMA using a combination of forward and backward recurrent networks consisting of 200 nodes. Thus, Bi-LSTMA considers both past and future input while evaluating the current hidden state, producing better information features [31]. Bi-directional LSTMs process the input data (CSI samples) in both directions to learn the context sequence of head orientations to make accurate recognition. CSI data is environment-dependent and consists of unimportant information induced by the surrounding objects. Bi-LSTM treats all inputs equally, making it difficult to filter out unimportant information from the sequence. It impacts recognition accuracy and increases the training cost. To address this, an attention layer of 400 nodes is added that assigns different weights to different time-step CSI sequences. The attention layer enables selective focus on the input sequence for generating output. It scores the feature vector, consisting of sequential inputs and output is connected to a dense layer of seven neurons, followed by a softmax to produce head gesticulation classification results. Adam optimizer is used to produce adaptive learning rates to optimize the parameters of the model. The loss function employed is the categorical entropy loss. The resulting CSI amplitude vector is transformed into a 10 × 10 CSI matrix using univariate-feature selection. A batch size of 128 is used while training the model. After recognizing head gesticulations, we now detect the intent of the speaker from the audio.

4.2 Intent Recognition from Audio

We recognize the speaker's intent using meeting audio. We first use the Google Speech Recognition API to extract the spoken utterance from the audio. Next, use a pre-trained intent recognition model *klue/roberta-small* [26] trained on 3*i*4*k* dataset [5] with 55134 samples to recognize the speaker's intention. The model returns seven intention labels such as *fragment*, *statement*, *questioning*, *command*, *rhetorical question*, *rhetorical command*, and *intonation-dependent utterance*. Since we are interested in correlating head gesticulations with the speaker's intent, we merge labels as per the speaker's context. For example, a participant's head gesticulations might be similar if it is a *questioning* vs *rhetorical question*. Hence, we group command & rhetorical command as *commands* and question & rhetorical question as *questions*. 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 [29]. We next correlate the listener's head gesticulation and the speaker's intent to monitor engagement level.

4.3 Monitoring Engagement

To compute the engagement score for the entire activity, we first divide it into 10s segments. First, we correlate the head gesticulations and speaker's intent for each segment. Next, we classify each segment as *engage* or *disengage* based on the correlation. Finally, we determine the engagement score for the entire activity after combining the segments. Details follow.


For each segment within the subcluster: Assign +1 to expected head gesticulations and -1 otherwise, as shown in table [1]. And then calculate the decision-metric D_i

Compute the maximum decision metric ➡ $D_{max} = max\{D_i, D_{i+1}, D_{i+2},..\}$.
If D_{max} > threshold, then classify each segment within the subcluster as engage (1); otherwise, disengage (0). In this way classify each segment as engage or disengage

Divide the entire meeting into smaller 10 second segments
S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8 S_{N-1} S_N
C_1
C_2
C_K
Cluster the segments based on listener's head gesticulation patterns
Re-cluster segments again based on speaker's intent
S_{11}
S_{15}
C_1E
C_1C
C_1Q
C_2Q
C_2E
C_KE
C_KC
C_KQ
E Explaining
C Command
Q Questioning
1 1 1 1 1 1 1 0 0 0

Figure 5: Classifying segment as engage(1)/disengage(0).

(a) Classification of segments: Fig. 5 shows the overall process. For each segment, we have multiple head gesticulations of the listener and one single intent. We scan all the meeting segments and create clusters of segments based on the similarity of the head gesticulation pattern. We use the Gestalt Pattern Matching algorithm [15] to identify similar head gesticulation patterns. We do not fix the number of clusters, which may vary from meeting to meeting. Now, for each cluster, we have a number of intents from multiple segments. For example, in a single cluster, some segments have *questioning* intent, and others have *explaining* intent. We further sub-cluster each cluster based on intent. For

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

Engage/Disengage	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 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

each segment within the sub-cluster, we compute a *decision-metric* $D_i = (\sum(+1 \times f_{en}) + (-1 \times f_{dis}))/f$. Where f_{en} is the total frequency of "expected" head gesticulations for that intent specified in Table. 1, f_{dis} is the total frequency of *not* "expected" head gesticulations for that intent and f is the total number of head gesticulations. Then, we compute the maximum decision metric (D_{max}) in that sub-cluster. Finally, each sub-cluster and its associated segments are classified as engage (1) if $D_{max} > 0.50$; otherwise, it is classified as disengage 0. We repeat the process for each sub-cluster. Finally, we obtain the engagement status (1/0) of each segment. Next, we compute the engagement score of the entire online activity.

(b) Computation of engagement score of entire online participation: We scan each segment one by one sequentially for all segments ($Segments_{total}$). We increment the value of En_{score} by 1 if the segment is classified as *engage*. However, if the model finds any segment classified as *disengage*, it does not decrement En_{score} 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 a longer duration might do. Hence, we check the next $n = 18$ segments. If the percentage of continuous *disengage* segments is greater than a threshold (80%), then we decrement En_{score} by the total number of *disengage* segments. We empirically find the optimal value of n and threshold. The final engagement score is computed as $En_{final} = \frac{En_{score}}{Segments_{total}}$.

5 USER STUDY AND DATA COLLECTION

To evaluate WiFiTuned, we ask the following research questions to analyze its performance under different aspects.

RQ1: How well can WiFiTuned classify each segment as engage or disengage and computes the engagement score of each participant for the entire meeting? 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.

RQ2: How well can WiFiTuned perform in the different locations with different meeting content? The engagement of the participants impacts by the meeting content and environment. The hypothesis is that WiFiTuned performs well in different locations with different meeting content.

RQ3: How well can WiFiTuned recognize the participant's head gesticulations using CSI modality? Pertaining to the difference in the underlying signatures for different head gesticulations, the hypothesis is that the head gesticulations recognizer accurately recognizes each head gesticulation.

5.1 Evaluation Methodology

Participants Details: We recruited 22 participants, $20 - 30$ years of age (10 female and 12 male), who attend online meetings frequently (once a day). The participants were well-informed about the tenure of the study, i.e. how long each meeting session will last, 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. While each of the participants was given a complete set of instructions regarding the study setup, 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. We collected the academic details of the participants, 12 are at the graduate level, 3 are at the postgraduate level, 3 are research assistants, and 4 are research scholars. Each participant was compensated with a 6.12 *USD* Amazon coupon.

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 select 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*.

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,

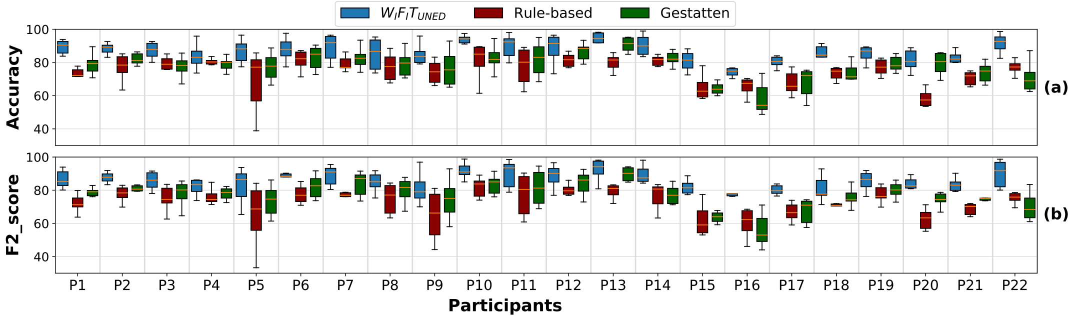


Figure 6: Participant-wise dist. (distribution) of (a) Accuracy and (b) F2_score at segment level.

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.

Setup: All participants join the online meetings using the *Zoom* platform with WiFiTuned setup. The setup has a laptop 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 (placed 1 − 2 meters apart) are flashed with CSI Tool Kit [13] 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 subcarriers). 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 feed are recorded from the laptop's mic and webcam, respectively.

5.2 Dataset Description

We collected a total of 3416708 time-stamped CSI samples from 132 meetings with a total duration of 42 hours and 35 minutes. The CSI data samples are labeled with *looking forward* (43.37%), *nodding/shaking*(18.00%), *looking down* (15.37%), *looking right and left* (13.4%), and *looking up* (10.12%). 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.

5.3 Ground Truth Generation

We split the frontal video feed of each participant into 10s video segments. We recruit 15 independent annotators to manually annotate the video segments as *engage* or *disengage*. We developed a website for annotation, where the annotators observe the participants in each video segment (randomly allocated) and annotate the video. Each video segment is annotated by three independent annotators. Whenever there is a disagreement (around 37% of the cases) among the annotators, we use a majority vote to mark a video segment as *engage* or *disengage*.

5.4 Baselines

We compare the performance of WiFiTuned with *Gestatten* (gaze gesture-based model) [16]. *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.

6.1 RQ1: Participant-wise Engagement Monitoring

As shown in Fig. 6, 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 behavioral 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 [12]: ***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.

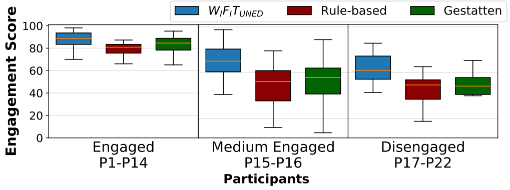


Figure 7: Dist. of the engagement scores at meeting level.

Fig. 7 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. 7, 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 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 most frequent was taking notes, texting, and using mobile. Gestatten and Rule-based fail to capture multitasking and hence incurs an average error of 16.33 and 17.82 respectively. ***(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.

6.2 RQ2: Meeting Content-wise & Location-wise Engagement Monitoring

6.2.1 ***Content-wise Engagement Analysis.*** Fig. 8 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 backward on the chair and leaning forward on the desk as they were sighing. 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.

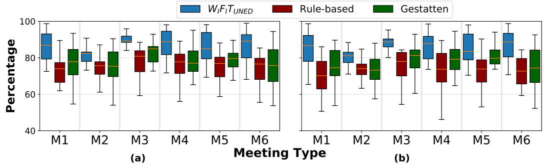


Figure 8: Content-wise dist. of (a) Accuracy, (b) F2_score at segment level.

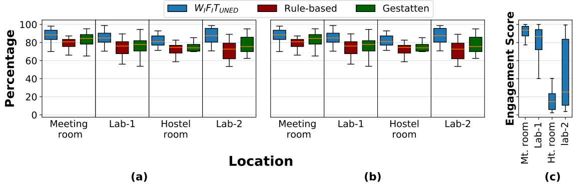


Figure 9: Location-wise dist. of (a) Accuracy, (b) F2_score at segment level, and (c) Engagement score at meeting level.

6.2.2 ***Location-wise Engagement Analysis.*** Fig. 9(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 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 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, lied 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 behavior 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 a situation 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. 9 (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 behavior of the participants.

6.3 RQ3: Head Gesticulations Recognition

We modeled this as a multiclass classification problem to recognize head gesticulations listed in Table 1. We train Bi-LSTMA and other similar state-of-the-art models such as *Constructive model*, *XGBoost (XGB)*, *Contrastive model*, *Support Vector Machine (SVM)*, *Gradient Boosting (GB)*, *Random Forest (RF)* and *Multiclass Logistic Regression (LR)* with collected CSI data. The performance of the recognition model is shown in Table 2. The Bi-LSTMA model provides higher performance scores (accuracy: 94%, precision: 93.57%, recall: 93.71%, F1-score: 94.57%) than other models. Bi-LSTMA, fed with time series input CSI data, uses the forward and backward context and better learns the temporal information of input sequence as is required to recognize head gesticulations (more details in §. 4.1.2). We further validate the performance of the recognition models using 10 fold cross-validation (each time, use nine fold as training and one fold as testing). The mean accuracy of Bi-LSTMA is 94.41% (better than others). Next, we trained the model using one participant out cross-validation, trained with 21 participants, and tested on 1. It achieves an average accuracy of 85.23% across all participants. Thus, we conclude the Bi-LSTMA 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 accuracy of 80%. Hence, the model is robust across four locations. The model recognizes each head gesticulation with an average accuracy of 93.75% (min 80.50%).

Table 2: Performance of Recognition Models

Model	Accu.	Precision	Recall	F1-score	Mean Acc.
Bi-LSTMA	**94.00%**	**93.57%**	**93.71%**	**93.57%**	**94.41%**
Contrastive	92.34%	89.36%	86.73%	87.92%	92.23%
XGB	92.23%	93.20%	92.33%	93.50%	63.17%
Constructive	57.25%	52.01%	49.28%	57.25%	52.35%
SVM	86.00%	82.14%	90.17%	85.17%	60.37%
GB	60%	43.25%	50.24%	52.35%	51.45%
RF	53%	27.85%	26.00%	25.14%	35.50%
LR	57%	29.57%	50.85%	32.14%	34.26%

6.4 Running Time

For a 10s segment, the preprocessing and denoising together take 0.5s, and extracting spoken utterances take 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.

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 ambiance plays a crucial role in determining the engagement level of participants. Interference from external sources can impact the CSI and thus WiFiTuned's performance. Further, participants are more prone to distractions in crowded indoor environments, and the presence of multiple users can impact the system's performance. To make WiFiTuned more robust and effective in diverse indoor environments, future research will focus on evaluating the system's performance in various indoor settings. This will help to identify and address any potential limitations or challenges that may arise in different environments and make the system more reliable and efficient in real-world scenarios.

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. It uses WiFi CSI to track and classify different types of users' head gesticulations. Further, the system utilizes the speech intent of the speakers and correlates it with the head gestures 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 extensively with multiple users under varied setups. The human studies reveal the significant efficacy of WiFiTuned and its promising applicability in promoting engagement in online meetings or providing attendees with insights into their engagement patterns.

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] Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E Mete, Eda Okur, Sidney K D'Mello, and Asli Arslan Esme. 2019. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *CHI 2019*. 1–12.

[3] 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.

[4] 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.

[5] Cheng Chang, Cheng Zhang, Lei Chen, and Yang Liu. 2018. An ensemble model using face and body tracking for engagement detection. In *ICMI 2018*. 616–622.

[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] 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.

[8] Betsy DiSalvo, Dheeraj Bandaru, Qiaosi Wang, Hong Li, and Thomas Plötz. 2022. Reading the Room: Automated, Momentary Assessment of Student Engagement in the Classroom: Are We There Yet? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–26.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] Amanda Lacy, Seth Polsley, Samantha Ray, and Tracy Hammond. 2022. A seat at the virtual table: Emergent inclusion in remote meetings. *Proceedings of the ACM on HCI* 6, CSCW2 (2022), 1–20.

[18] 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.

[19] 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.

[20] 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.

[21] Arien Mack. 2003. Inattentional blindness: Looking without seeing. *Current directions in psychological science* 12, 5 (2003), 180–184.

[22] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, Paul Johns, and Akane Sano. 2016. Neurotics can't focus: An in situ study of online multitasking in the workplace. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1739–1744.

[23] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cécile Paris. 2020. Automatic recognition of student engagement using deep learning and facial expression. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 273–289.

[24] 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.

[25] 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.

[26] 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]

[27] 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.

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

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

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

[31] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. The Performance of LSTM and BiLSTM in Forecasting Time Series. In *2019 IEEE International Conference on Big Data (Big Data)*. 3285–3292.

[32] 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.

[33] Mohamed Soltani, Hafed Zarzour, and Mohamed Chaouki Babahenini. 2018. Facial emotion detection in massive open online courses. In *World Conference on Information Systems and Technologies*. Springer, 277–286.

[34] Wei Sun, Yunzhi Li, Feng Tian, Xiangmin Fan, and Hongan Wang. 2019. How Presenters Perceive and React to Audience Flow Prediction In-situ: An Explorative Study of Live Online Lectures. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–19.

[35] 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.

[36] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *MobiCom 2015*. 65–76.

[37] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2017. Device-free human activity recognition using commercial WiFi devices. *IEEE Journal on Selected Areas in Communications* 35, 5 (2017), 1118–1131.

[38] Zhengyang Wang, Sheng Chen, Wei Yang, and Yang Xu. 2021. Environment-Independent Wi-Fi Human Activity Recognition with Adversarial Network. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3330–3334.

[39] 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.

[40] 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.

[41] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *MobiSys 2019*. 313–325.