SEMPI: A Database for Understanding Social Engagement in Video-Mediated Multiparty Interaction

Maksim Siniukov*
Institute for Creative Technologies,
University of Southern California
Los Angeles, California, USA
siniukov@usc.edu

Yufeng Yin*
Institute for Creative Technologies,
University of Southern California
Los Angeles, California, USA
yufengy@usc.edu

Eli Fast
El Camino College
Torrance, California, USA
eclfast@gmail.com

Yingshan Qi
University of Southern California
Los Angeles, California, USA
qiyingsh@usc.edu

Aarav Monga
University of Southern California
Los Angeles, California, USA
armonga@usc.edu

Audrey Kim
University of Southern California
Los Angeles, California, USA
amkim@usc.edu

Mohammad Soleymani
Institute for Creative Technologies,
University of Southern California
Los Angeles, California, USA
soleymani@ict.usc.edu

ABSTRACT

We present a database for automatic understanding of Social Engagement in MultiParty Interaction (SEMPI). Social engagement is an important social signal characterizing the level of participation of an interlocutor in a conversation. Social engagement involves maintaining attention and establishing connection and rapport. Machine understanding of social engagement can enable an autonomous agent to better understand the state of human participation and involvement to select optimal actions in human-machine social interaction. Recently, video-mediated interaction platforms, *e.g.*, Zoom, have become very popular. The ease of use and increased accessibility of video calls have made them a preferred medium for multiparty conversations, including support groups and group therapy sessions. To create this dataset, we first collected a set of publicly available video calls posted on YouTube. We then segmented the videos by speech turn and cropped the videos to generate single-participant videos. We developed a questionnaire for assessing the level of social engagement by listeners in a conversation probing the relevant nonverbal behaviors for social engagement, including back-channeling, gaze, and expressions. We used Prolific, a crowd-sourcing platform, to annotate 3,505 videos of 76 listeners by three people, reaching a moderate to high inter-rater agreement of 0.693. This resulted in a database with aggregated engagement scores from the annotators. We developed a baseline multimodal pipeline using the state-of-the-art pre-trained models to track the level of engagement achieving the CCC score of 0.454. The results demonstrate the utility of the database for future applications in video-mediated human-machine interaction and human-human social skill assessment. Our dataset and code are available at https://github.com/ihp-lab/SEMPI.

KEYWORDS

Engagement, Multiparty Interaction, Dataset, Machine Learning

ACM Reference Format:
Maksim Siniukov, Yufeng Yin, Eli Fast, Yingshan Qi, Aarav Monga, Audrey Kim, and Mohammad Soleymani. 2024. SEMPI: A Database for Understanding Social Engagement in Video-Mediated Multiparty Interaction. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685752

1 INTRODUCTION

Social engagement is the degree of attention and involvement of an interlocutor in a conversation. Understanding the attentiveness and participation of listeners in social interaction can provide a valuable social signal for the measurement of rapport, active listening, and interaction quality. Machine understanding of engagement can enable a multitude of applications. An agent facilitator which can sense the engagement and involvement of its social counterparts can re-calibrate its conversational strategies to improve interaction quality. Machines could also be used to assess human-human interaction in soft skill training and evaluation [11, 20, 48]. With the growing popularity of video-mediated communication platforms, *e.g.*, Zoom, our social interactions, including support groups and group therapy sessions, are moving online.

Despite the increasing prevalence of video-mediated interaction, to the best of our knowledge, there is no database of video-mediated social interactions with emotional and social labels. There are also

*equal contribution

ICMI '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685752

only a few labeled databases for understanding social engagement in dyadic and multiparty interactions between humans [7, 45] and humans and robots [3, 10, 31] interactions. However, no database has been developed with the goal of understanding social engagement in multiparty human-human video-mediated interactions. Moreover, none of the publicly available databases are related to the domain of mental health and support groups.

In this work, we present a database for understanding the Social Engagement of listeners in video-mediated MultiParty Interaction (SEMPI). Leveraging publicly available recorded virtual meetings primarily consisting of pre-recorded support groups on YouTube, we develop a database of listener videos annotated for their level of engagement. We have created this database to help us train tools that can measure participation and group climate in support groups and group therapies. To improve the consistency and validity of the annotations, we designed a rating scale with six questions that can capture different behaviors that are associated with engagement and disengagement, *e.g.*, the presence of back-channeling and distracted behavior. After segmenting video calls by utterance, cropping, and filtering the videos that are of low quality or low duration, we annotated 3,505 videos of 76 listeners' reactions in multiparty video calls through crowdsourcing. The inter-rater agreement among the raters is 0.693 measured by Krippendorff's alpha [29]. We then train a multimodal machine learning model for detecting the level of engagement, achieving the CCC score of 0.454.

This database can be used to analyze nonverbal and verbal behaviors of interlocutors in video-mediated multiparty interactions. This database is particularly useful for developing models that can track listener engagement to measure interaction quality and group connectedness in video-mediated multiparty interactions. Given the common usage of virtual meetings in mental healthcare, this database is particularly useful for research on empathy and social cues in such settings. The major contributions of this work are as follows.

- We introduce a publicly available database of speaker and listener videos in multiparty video calls with annotations regarding their engagement, distraction, and the frequency of back channels.
- We develop a multiscale questionnaire to assess social engagement. This questionnaire results in superior inter-rater agreement compared to a simple rating of engagement, reaching a moderate to high inter-rater agreement of $\alpha = 0.693$.
- We propose and evaluate a baseline multimodal model for detecting the level of listener engagement, achieving the CCC score of 0.454 in a participant and video call independent cross-validation. The results demonstrate the utility of the proposed database.

2 RELATED WORK

2.1 Engagement

Engagement is a complex concept studied in various contexts [21, 40, 46]. Engagement can be described as an affective, cognitive, or behavioral construct [16]. The cognitive view of engagement focuses on conscious components or effort. The affective view of engagement considers engagement an emotional state [44]. Finally, the behavioral view considers actions and participation to characterize engagement. Engagement in interaction is defined as maintaining joint attention, performing coordinated activities, and establishing connection and rapport.

Engagement is related to motivation, interest, experience, attention, and action [40]. Rich *et al.* [44] identified four cues for measuring engagement, including gesture, speech, gaze, and conversational activity.

This paper focuses on social engagement that includes the processes by which people establish and maintain their connection, including verbal responses and nonverbal acknowledgments, *e.g.*, head nods [49]. There is a wide body of work on understanding engagement with tasks or educational material [14, 27]. Student engagement is different from social engagement as it mainly concerns task engagement and is related to mind wandering and focus.

2.2 Engagement Detection

Behavioral markers associated with engagement can be tracked through audiovisual verbal and nonverbal behavior analysis. Methods for recognizing engagement are similar to those used in emotion recognition and rely on identifying patterns in behavioral and physiological changes.

Engagement recognition methods may be divided into two broad categories: traditional machine learning approaches with handcrafted features and deep learning approaches. Traditional machine learning methods often rely on handcrafted features inspired by social psychology, which often include visual cues such as gaze [35, 47] and facial action units (AUs) [15, 22, 55]. Interpersonal cues often serve as reliable features for engagement detection. Proximity or the relative distance between people in a conversation has been used for tracking engagement [47]. Body behaviors have been also used for estimating engagement, *e.g.*, body posture [26, 50], and global quantity of movement during interaction [47, 50] that may be computed from skeleton joints in both Human-Robot and Human-Human Interactions (HRI and HHI). Linguistic cues such as greetings may be a sign of engagement [4]. Spoken language has been used for detecting engagement [19, 59]. Vocal features can be also used to detect engagement, albeit more for the speakers. Oertel *et al.* [37] used basic vocal features – voice span and intensity – to detect involvement in conversation.

A variety of classical machine learning models has been applied to address engagement estimation task, including support vector machine (SVM) [9, 18, 47, 55, 59], linear regression [18, 22], Hidden Markov Models [25, 26], Naïve Bayes [12, 25], and Ensemble-based methods [5, 18, 25, 47]. Deep learning approaches have shown superior performance for engagement estimation tasks, including convolutional neural networks (CNN), recurrent neural networks (RNN) [12, 15, 25, 51], multi-layer perceptron (MLP) [18, 25] and multimodal transformer models [51].

2.3 Databases for Engagement Understanding

In this section, we review the most relevant publicly available databases (see Table 1). Databases with engagement scores include human-human and human-robot interactions.

Table 1: Overview of the publicly available relevant engagement and multimodal interaction databases. # S denotes the number of subjects. TL means the total length. Biophysical data refers to various signals such as temperature, electrodermal activity, wrist acceleration, and EGG.

Database	# S	TL	Modality	Annotations	Interaction Type
MHHRI [10, 33]	18	4h	audio, video, biophysical	personality, self-reported engagement	HRI / HHI
RECOLA [45]	46	4h	audio, video, EDA, ECG	ordinal (7-point scale) ratings of engagement	HHI
NoXi [7]	87	25h	audio, video, depth, body skeleton	discrete/continuous (behavioral, engagement)	HHI
UE-HRI [3]	54	-	audio, video, sonar, laser, depth	ordinal engagement labels	HRI
PInSoRo [31]	120	45h	audio, video and depth	ordinal engagement labels	HRI/HHI
SEMPI (ours)	76	7h30m	audio, video, transcript	continuous engagement scores and five other relevant cues	HHI

The Remote Collaborative and Affective Interactions (RECOLA) [45] includes remote dyadic human-human spontaneous interactions. The database provides continuous emotion annotations as well as engagement scores. However, the engagement labels were obtained from the post-study questionnaire by asking the person whether they enjoyed the interaction. The expert NOvice eXpert Interaction (NoXi) database is a multilingual human-human dyadic database [7], where all lab-based interactions were performed over video. The database includes audio, visual, depth, body movements, and face features data with additional social signals markup. Manually annotated behavioral cues include low-level social signals such as gestures and smiles, functional descriptors such as turn-taking and dialogue acts as well as interaction descriptors such as interest, fluidity, and discrete engagement labels.

Given the significance of engagement in human-robot interaction, a number of HRI databases have been developed. Multimodal Human-Human-Robot Interactions (MHHRI) database [10] is a multimodal database for studying how personality affects engagement in human-human and human-robot interaction. The dataset was collected in a controlled environment, where people participated in two settings: dyadic interactions between two human participants and interactions between two human participants and a robot. In both setups, participants were asked personal questions. The engagement labels were obtained via a post-study questionnaire by asking people whether they enjoyed the interaction. Later study [47] used the Temple Presence Inventory (TPI) questionnaire [33] to collect additional labels for the MHHRI database, where external observers annotated engagement labels via an online crowd-sourcing platform. Another database focusing on spontaneous social HRI is the User Engagement in Spontaneous HRI (UE-HRI) database [3]. Participants interacted with a humanoid robot, *i.e.*, Pepper, for four to 15 minutes. The database was collected using a variety of sensors, including directional microphones, cameras, and depth sensors. User feedback was captured through the robot's touch screen. Annotators labeled the collected data by first finding the beginning and the end of each interaction and then by indicating engagement breakdown and temporary disengagement and assigning corresponding negative effects for the segment (*e.g.*, nervousness, boredom). PInSoRo [31] is a database of child-child and child-robot interaction with children engaging in virtual free-play with a sandbox accompanied by a humanoid Nao robot. Social and task engagement was annotated by coders, indicating the level of cooperation and involvement of children with each other, the robot, and the game. More than 45 hours of videos were collected from 120 children.

None of the existing databases addresses listener engagement in multiparty human-human interaction in-the-wild. Even though NoXi [7] and RECOLA [45] are video-mediated, they are both recorded in the lab and are dyadic interactions. CANDOR [43] is a large open dyadic database without engagement annotations. Its license is restrictive; therefore, we could not annotate it for engagement to create a new database, as no derivatives are allowed. Hence, SEMPI fills this gap by providing listener-focused engagement ratings with in-the-wild videos. Technical details of the datasets, including the number of subjects, total session duration, modalities used, setting, and interaction type, are described in Table 1.

3 SEMPI DATABASE

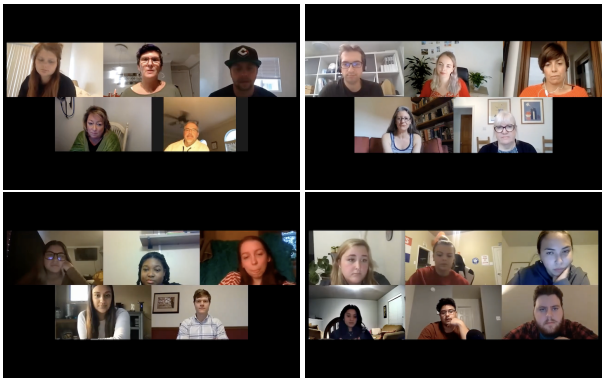


Figure 1: Screenshots of the selected videos from YouTube for social engagement understanding.

Table 2: Details of the videos for engagement understanding. 14 videos are collected from YouTube in which video ID (vid) are provided. # S denotes the number of subjects.

YouTube vid	Start time	End time	Length	# S
E5rDlGZr-bM	00:00:00	01:05:00	1 h 06 min	6
6jY61ZTbzFw	00:00:00	00:45:00	45 min	9
Ej1qN1mTU4g	00:14:22	01:07:57	54 min	5
dQWDgDdmk9s	00:56:42	01:31:32	35 min	5
LLpot3VFXmA	00:05:57	00:25:00	20 min	4
CtZMLcelacA	00:00:00	00:51:35	51 min	5
kFPSpoGmYaE	00:00:00	00:15:48	16 min	5
YO1ntZ23uRg	01:03:02	01:38:30	35 min	5
qQqsjvDubSo	01:03:00	01:31:00	28 min	7
u1QMSLUIqNg	00:00:00	00:24:21	24 min	3
InI2NK_5fps	00:00:00	00:45:30	45 min	4
bT1NIyVJGQw	01:47:00	02:22:00	35 min	7
L273NSU1hhw	00:00:00	00:40:33	41 min	6
a6kAdamVXIE	00:04:12	00:35:24	31 min	5
Total	-	-	8 h 46 min	76

3.1 Overview

In this work, we present the SEMPI database for social engagement understanding in video-mediated multiparty interaction. The dataset consists of in-the-wild videos of virtual meetings from YouTube (see Table 2 and Figure 1). Except for one meeting, all videos are sourced from pre-recorded support groups that are posted on YouTube. The development and annotation of this database have been reviewed and approved by the University of Southern California's institutional review board (IRB). Content owners have been informed about the study and agreed or did not object to the usage of their data for research. To ensure the label quality and high agreement among raters, we propose to annotate the listeners' engagement based on multiple dimensions that can be combined to generate an aggregate engagement score (refer to Table 3), which yields an inter-rater agreement of **0.693** by Krippendorff's alpha.

Overall, the database comprises listener responses to 3,505 utterances with around 7.5 hours of audiovisual content from 76 people. Each utterance is annotated with an engagement score ranging from −1 to 1. The database also includes the speakers' videos. However, the speaker videos are not annotated for engagement since the annotation scheme was designed for listeners. Each sample is five to ten seconds long with a cropped video feed, audio, and the corresponding transcript. Examples of the database are shown in Figure 2.

3.2 Data Collection and Processing

Data collection. To understand social engagement in multiparty interaction, we collect in-the-wild videos of virtual meetings from YouTube. Specifically, we select videos from various YouTube channels according to the following requirements: (1) Relevance to support group topics, ensuring each participant interacts with the host and other speakers. (2) A maximum of nine speakers per video, with a minimum resolution of 720p, to ensure clear video quality

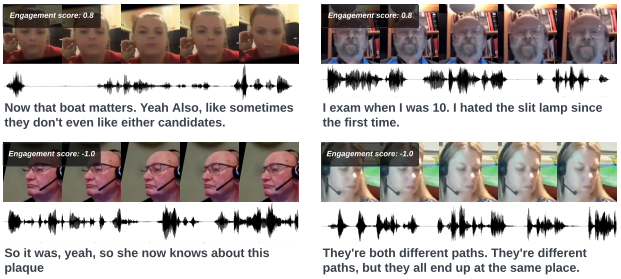


Figure 2: Examples of high engagement (first row) and low engagement (second row). Each sample has a cropped face, audio, and transcript. We annotate the listeners' engagement score ranging from -1 to 1.

suitable for engagement analysis. (3) A minimum video length of 30 minutes to provide sufficient data for model training and evaluation. (4) Sourcing data from a variety of channels to ensure a diverse database.

Overall, we collected 14 videos of virtual meetings with 76 people, which was about nine hours long in total. To simplify the processing pipeline (cropping and segmenting), for every video, we selected the period where the number of speakers and their corresponding positions in the videos are fixed. Detailed information for each video is provided in Table 2. Screenshots of selected videos are shown in Figure 1.

Data processing. We show the overview of the data processing pipeline in Figure 3. Given a raw video with multiparty interactions, we first use voice activity detection from PyAnnote [6, 41] to get the timestamps of each speech activity and then segment the video into utterances. Each utterance is about five to ten seconds long and contains only one speaker's speech. Subsequently, we crop the whole frame into images of individual speakers. We then use Whisper-tiny [42] for speech recognition to generate the transcripts. We utilize dlib [28] for facial landmark detection to crop and align the faces. Following this procedure, we obtain an initial set of 6,400 video clips with cropped faces, audio, and the corresponding transcript.

3.3 Questionnaire Design

Past work, *e.g.*, Noxi [7], simply asked the raters to indicate the level of engagement by the speakers on a multi-point scale. Such an annotation process may result in a low agreement among multiple coders, as we initially observed. Therefore, we propose to annotate the listeners' engagement based on multiple dimensions. In particular, we had two coders watch a small set of videos and write down observations of what behaviors constitute low or high engagement. The observations are then aggregated into five questions with non-overlapping markers (see Q1 to Q5 in Table 3, step 3). The annotators are asked to answer each question on a five-point scale (1-never, 2-rarely, 3-sometimes, 4-often, 5-very often). Finally, they will be asked about the listeners' overall engagement (see Q6 in Table 3, step 3). Q6 has three answer choices (yes, unsure, no). We have decided to use a different scale for the last question, as the coders have found directly distinguishing engagement beyond three levels difficult.

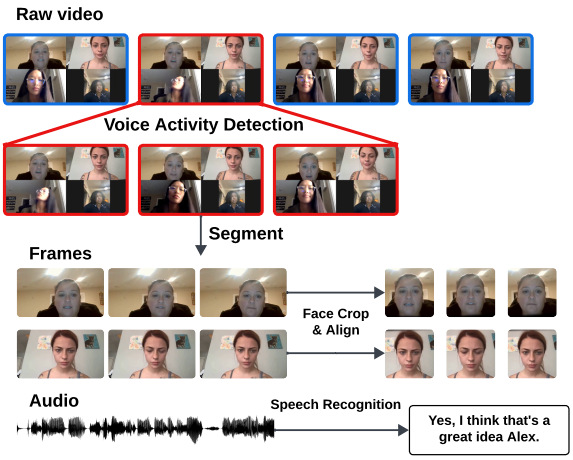


Figure 3: Overview of the data processing pipeline. (1) We first use voice activity detection to segment the raw video into utterances. (2) We crop the whole frame into video sequences of individual speakers. (3) We crop and align the faces and generate the transcript with speech recognition.

3.4 Data Annotation

The questionnaire for the engagement annotation was delivered through *Qualtrics*, a popular platform used to create surveys and gather responses. We used *Prolific*, an online platform designed to connect researchers with participants for academic studies and market research for all our annotations. We only recruited English-speaking participants from the United States. The crowd-workers were compensated on average with an effective average hourly rate of $12 per hour.

To maintain consistent annotation standards among participants, we provided annotation guidance at the beginning of the survey, including examples of high and low ratings for each question (see Table 3, step 1). Additionally, to reduce annotation confusion, we introduced three filtering questions before presenting the main questions for every video clip to be annotated. Specifically, data samples falling into the following three categories were skipped for annotations: (i) No one is visible in the video, or the camera is turned off. (ii) The individual in the video is speaking rather than listening. (iii) Low video quality. (refer to Table 3, step 2). Examples of the skipped video clips are shown in Figure 4.

To ensure the label quality, we first conducted a pilot study with five sample videos to identify the best annotators. Participants are assumed to be qualified if they achieve a high level of agreement with our internal annotations. In particular, we screened 300 annotators for the pilot study, from which 170 were invited to the main annotation phase.

In the main annotation phase, each annotation batch contained 20 videos, which were randomly selected from the whole pool of video clips. The video clips are not selected if they have already been annotated three times. The random sampling of videos and their annotation status on Qualtrics was managed by a backend server developed in Python. Overall, the selected 170 coders annotated 960 batches of 20 video batches. The average annotation time for each

Table 3: Proposed questionnaire for engagement annotation. During the annotation process, (1) annotation guidance is shown to maintain consistent annotation standards among participants; (2) three filtering questions are introduced to reduce annotation ambiguity; and (3) the listeners' engagement is annotated on multiple dimensions, the responses to six questions are aggregated to yield a high inter-rater agreement.

Step 1. Annotation Guidance You will watch a short video of a person participating in a Zoom meeting. You will then be asked six questions to evaluate the person's engagement. If no one is in the video or the camera is turned off, you may skip the annotation questions. If the person is the main speaker or the video quality is too poor to evaluate the person's engagement in the meeting, you may also skip the annotation questions. The main speaker is the person who is presenting for the majority of the meeting clip; speaking for 1-2 seconds does not count. Please answer these questions truthfully; your answers will be checked for accuracy, and if they are found incorrect, you may not be compensated for this survey. For Q1 to Q5, answer the questions on a scale of 1-5 (1-never, 2-rarely, 3-sometimes, 4-often, 5-very often). Q6 has three answer choices (yes, unsure, no). Please answer these questions to the best of your ability. Examples of each question with high and low ratings are presented in the survey.
Step 2. Filtering Questions **Q1**: No one is in the video, or the camera is turned off. **Q2**: Is the person in the video the main speaker? (The main speaker is the person who is presenting for the majority of the meeting clip; speaking for 1-2 seconds does not count.) **Q3**: Is the video quality too poor to see the person's features clearly? If any one of the answers is true, step 4 will be skipped.
Step 3. Main Questions **Q1**: The listener **produced verbal/audible sounds** in response to the conversation. (*e.g.*, As the speaker was presenting, the listener said "yeah" and "uh huh") **Q2**: The listener **nodded or shook their head** in response to the conversation. (*e.g.*, As the speaker was presenting, the listener was nodding their head in agreement) **Q3**: The listener made **facial expressions** in response to the conversation. (*e.g.*, The speaker was telling a sad story, and the listener frowned with their eyebrows furrowed) **Q4**: The listener was **engaged in some other activity** other than listening to the speaker. (*e.g.*, The listener was typing while the speaker presented.) **Q5**: The listener **looked away** from the speaker. (*e.g.*, The listener was looking elsewhere and not at the speaker) **Q6**: Overall, would you say that the listener was **socially engaged** with the speaker? (*e.g.*, The listener was paying attention to words and tones while considering the emotion and perspective of the speaker.)

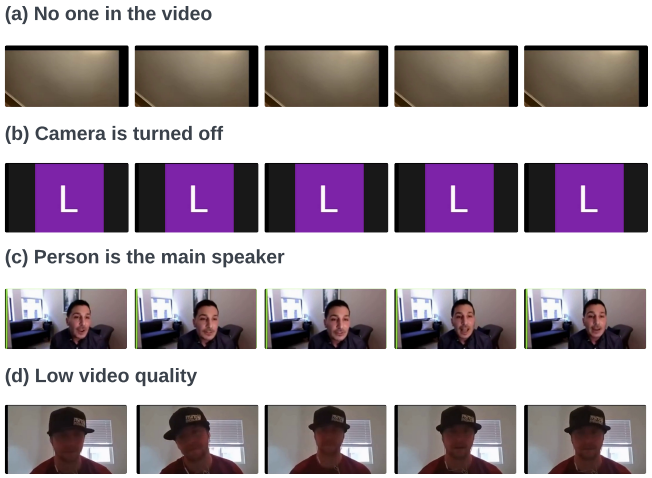


Figure 4: Examples of the skipped video clips for engagement annotations.

video clip was around one minute and 20 minutes for the whole batch.

3.5 Post-processing

To calculate the overall engagement score, we collect six scores for each utterance, representing responses to the main questions. Formally, we denote the answers as a_i, where $i \in [1, 2, ..., 6]$. The last question offers three choices, which we convert using the following mapping: 1 for "no," 3 for "unsure," and 5 for "yes." Subsequently, all the six responses range from 1 to 5. Q1, Q2, Q3, and Q6 contribute to high engagement, while Q4 and Q5 are related to low engagement. We aggregate the answers and normalize the result to $[-1, 1]$.

$$e = (a_1 + a_2 + a_3 - a_4 - a_5 + a_6 - 6)/12, \quad (1)$$

where e is the overall engagement score.

Out of 6,400 utterances, we have 4,705 annotated at least twice and 3,505 annotated at least three times. Some utterances are skipped for annotation by crowdworkers due to the filtering questions, e.g., whether the video feed shows a face. The final SEMPI database comprises 3,505 labeled utterances, in addition to the corresponding unlabeled speaker videos. For each utterance, we calculate the final engagement annotation by averaging the responses from the participants. We measure the inter-rater agreement by Krippendorff's alpha [29]. The agreement score for utterances annotated twice is **0.688**, and for those annotated three times is **0.693**. The results indicate a moderate to high level of agreement among raters, demonstrating the utility of our proposed annotation process of engagement.

3.6 Data Observations and Analysis

We analyze the engagement annotations as well as the statistics for the SEMPI database. Specifically, we present the histogram illustrating the distribution of engagement scores in Figure 5a. We observe that the distribution closely resembles a bell curve with a slight skewness towards positive engagement and only a minority of utterances displaying notably high or low engagement scores. More than half of the data samples fall within the range of values near 0, ranging from −0.25 to 0.25. The highest recorded engagement score is 0.83, while the lowest is −1.0.

Additionally, we count the number of utterances extracted per video meeting and per subject in descending order in Figures 5b and 5c, respectively. The video meetings with the highest and lowest utterance counts contain 531 and 57 samples, contributing to 15.1% and 1.6% of the database, respectively. Six out of 14 videos have more than 200 utterances. The subjects with the highest and lowest utterance counts contain 112 and 1 sample, contributing to 3.2% and 0.03% of the database, respectively. 35 out of 76 subjects have more than 40 utterances. The observations suggest that our database is diverse, and each video and subject has a sufficient number of samples for training and evaluation of machine learning models for understanding social engagement.

4 METHOD

To demonstrate the validity and utility of the developed database, we leveraged it to train and evaluate a multimodal machine learning model for estimating social engagement.

In this section, we introduce a baseline multimodal model for engagement estimation (see Figure 6). The method incorporates multimodal information, including vision, language, and speech. A set of features were selected, based on their past usage in a similar context [36, 58]. For the visual modality, we utilize both raw frames and high-level visual features, *e.g.*, facial action units [17], facial landmarks, head pose, and gaze. We employ separate encoders to extract meaningful representations from each modality. Subsequently, features from different modalities are concatenated and fed into a linear regression model to predict the engagement score.

4.1 Problem Formulation

Engagement estimation. Given a participant's video clip with the corresponding audio and transcript, we aim to estimate their engagement score on a continuous range from −1 (no engagement) to 1 (fully engaged) using function F.

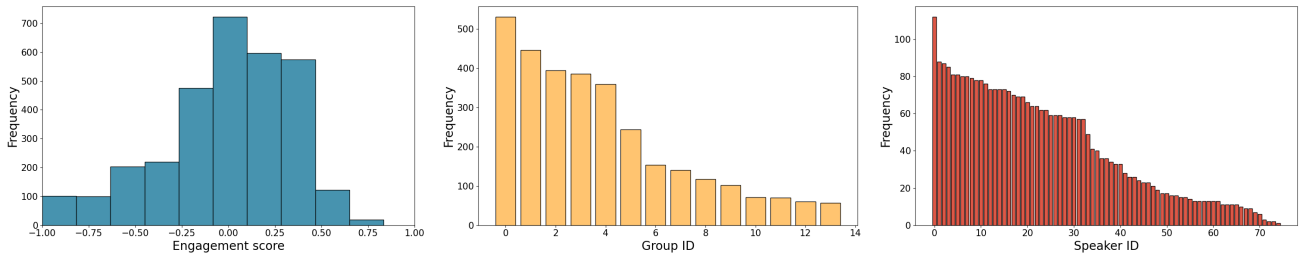
$$e = \mathrm{F}(v, a, t), \quad (2)$$

where e denotes the engagement score, v, a, and t refer to the video, audio, and text, respectively. One noteworthy point is that we focus on studying the listeners' engagement; thus, we utilize the listeners' videos as input. Given that listeners typically do not vocalize much, we utilize the speakers' audio and text as inputs for the engagement estimation model.

4.2 Feature Extraction

Visual features. We extract deep visual features using the Inception_I3d model [8, 58], which extends the traditional Inception model [52] to video via 3D convolutions. Specifically, the model captures spatio-temporal features, resulting in a 1024-dimensional feature vector h_{v1}.

High-level visual features. Following prior work [36], we extract both facial and head-pose features for each frame with OpenFace 2.0 [53], which is an open-source framework for facial behavior analysis. In particular, we extract 35 facial action unit intensity and



(a) Distribution of engagement scores. **(b) Number of utterances per group.** **(c) Number of utterances per subject.**

Figure 5: Data observation for the SEMPI database. (a) The histogram of the engagement scores shows that most engagement scores are concentrated in the middle, with a slight asymmetry towards positive engagement. (b) Videos were extracted from different video calls, with some calls contributing fewer samples than others. (c) Given that the majority of samples are from a few video calls and some participants appear in different calls, the distribution of instances per participant is not uniform, with a single participant having more than 100 samples.

presence [1], six head pose features [2], and eight gaze features [56], 56 3D facial landmarks, and 56 2D facial landmarks [60], resulting in a total of 329 features. These high-level features are mean-pooled across the temporal dimension and then encoded into a more dense 64-dimensional latent representation h_{v2} via a linear projection layer.

Speech features. Each video clip is accompanied by a corresponding audio recording that captures the speaker's speech. We extract audio features using the HuBERT-base model [23], which is pre-trained with masked modeling of hidden units objective. The HuBERT model shares a similar architecture to BERT [13], and it has demonstrated high performance in various downstream tasks for signal processing, including continuous emotion recognition [54] and keyword spotting [57]. The speech features from the last layer are mean-pooled across the temporal dimension, resulting in a 768-dimensional vector h_a per audio sequence.

Language features. We utilize the transcripts generated from the speaker's audio and feed them to a pre-trained language model, *i.e.*, RoBERTa [32]. RoBERTa is an encoder-only model, similar to BERT [13], trained with a masked language modeling objective on a large corpus of text. The model yields semantically rich hidden states, which show competitive performance on various downstream tasks such as text classification [32]. We extract the text features from the last layer of the encoder and employ mean pooling to aggregate the temporal dimension, resulting in a 768-dim feature h_t.

4.3 Multimodal Fusion

With the encoded hidden representations h_{v1}, h_{v2}, h_a, and h_t, we concatenate them together and input the concatenated features into a multilayer perceptron (MLP) (three fully-connected layers) to estimate the person's engagement level.

$$h = \text{Concat}(h_{v1}, h_{v2}, h_a, h_t), \tag{3}$$

$$\hat{e} = \text{MLP}(h). \tag{4}$$

where $\hat{e}$ is the estimated engagement score.

4.4 Training Objective

The learning objective for the baseline method is the Concordance Correlation Coefficient (CCC) loss between the predictions and the ground-truth labels.

$$\mathcal{L}_{\text{CCC}} = 1 - \text{CCC}(\hat{e}, e), \tag{5}$$

$$\text{CCC}(x, y) = \frac{2 s_{xy}}{s_x^2 + s_y^2 + (\overline{x} - \overline{y})^2}, \tag{6}$$

where $\hat{e}$ and e are the predictions and ground-truth labels respectively.

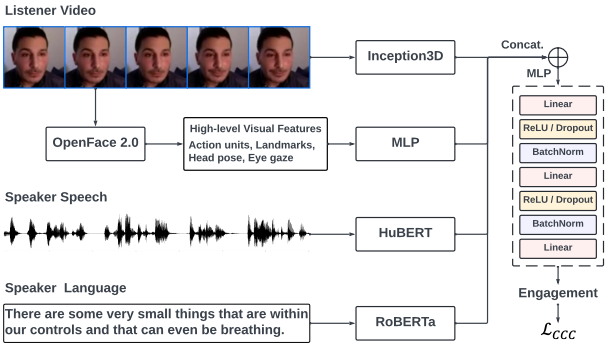


Figure 6: Overview of the multimodal baseline for engagement estimation. The model inputs include vision (raw frames and high-level visual features), audio, and language modalities. Each modality is fed into a separate encoder, and their outputs (embeddings) are fused to predict the engagement score.

5 EXPERIMENTS

5.1 Implementation Details

All methods are implemented in PyTorch [38]. All models were trained on a single NVIDIA L40S GPU.

Model architecture. To avoid over-fitting and make training efficient, we only fine-tuned the last two layers for the visual, speech, and language encoders. The MLP for engagement estimation comprises three fully connected layers where the number of hidden units is 32. Each linear layer is followed by a batch normalization layer [24]. A 10% dropout is applied to linear layers.

Model training. We optimized the network weights using the AdamW optimizer [34] with a learning rate of 3e-4. The batch size was set to 8. We applied regularization to avoid over-fitting via a weight decay with a 0.01 penalty coefficient. We trained the models for 30 epochs with an early stopping technique — if the validation performance did not improve for ten consecutive epochs, the training was stopped.

5.2 Evaluation metrics

Two evaluation metrics are used to assess the performance of the continuous engagement estimation, *i.e.*, Concordance Correlation Coefficient (CCC) [30] and Pearson Correlation Coefficient (PCC) [39]. These two metrics measure the agreement between the predictions and reference engagement values. The CCC and PCC coefficients range from -1 for perfect disagreement to 1 for perfect agreement. Higher CCC and PCC refer to better model performance.

Table 4: Performance of engagement estimation on SEMPI database. CCC (↑) and PCC (↑) are reported for participant-independent five-fold cross-validation. Numbers in the brackets are standard deviations measured across folds. "-" indicates removing a modality.

Method	CCC	PCC
All features	0.454 (0.110)	0.496 (0.100)
- speech	0.430 (0.102)	0.477 (0.099)
- language	0.401 (0.115)	0.453 (0.104)
- speech - language	0.394 (0.127)	0.413 (0.131)
- high-level visual	0.418 (0.115)	0.453 (0.104)
- visual	0.188 (0.108)	0.212 (0.113)
- high-level visual - visual	0.102 (0.045)	0.127 (0.054)

5.3 Experimental Results

We used subject-video-independent five-fold cross-validation to evaluate the performance of engagement estimation. Specifically, we allocated 80% of videos to the training set and 20% to the validation set. Data samples from the same video meeting or subject did not appear in both training and validation data.

We report the performance of engagement estimation in Table 4. The multimodal model achieves a CCC score of 0.454 and a PCC value of 0.496. In an ablation study, we analyzed the importance of each modality on engagement prediction by removing each unimodal encoder from the model. We observe that similar to the past work, visual modality is the most important channel of information, with a CCC score of 0.394. This is superior to the results reported for facial expressions in NoXi [36] (CCC= 0.31). Unlike NoXi [36], we do not have a full video of the upper body in these in-the-wild videos, hence, body pose features cannot be used with this dataset. Speech and language, even though they mostly come from the speaker rather than the listeners, contribute to the estimation. Speech was not diarized, and it is possible that the vocalized backchannels could contribute to this. However, given that this is a multiparty interaction and the effect of the paralinguistic vocal features according to the ablation is minimal, we speculate that the effect is negligible for an individual listener. This shows that the variations in the verbal and nonverbal features in speech affect the engagement of the listeners (CCC of 0.188, which is more than the chance level of 0). The multimodal fusion achieves the best result, combining all the features from the speaker's speech and the listener's visual behaviors.

6 CONCLUSIONS

In this paper, we present SEMPI, a novel database for understanding social engagement in multiparty video-mediated interactions. The database primarily consists of participant videos in online support groups. Hence, the database can serve as a research resource for understanding support groups and group therapies. To annotate the level of engagement of listeners, we developed a multi-scale questionnaire that assesses typical listener behaviors associated with active and engaged listening, *e.g.*, back-channeling. We segmented and cropped videos of video calls into individual videos of responses to utterances of five to ten seconds long. The videos were annotated by two to three raters on Prolific and the ratings were aggregated to form a continuous engagement score. We trained a multimodal machine learning model to detect engagement, achieving a CCC score of 0.454. The result achieved by a relatively simple model demonstrates the utility of the proposed database for automatic engagement understanding. This database can serve as a resource for further research on the behavioral markers of engagement and group climate in multiparty interaction in the mental health context.

Limitations and future work. Our approach relies on third-person annotations, focusing on observed engagement rather than first-party annotations. Our primary goal is to build an agent who tracks observable engagement, similar to humans, in multiparty interactions for group facilitation and management, however, the dataset does not capture the subjective experience of the participants. In addition, we could not reach all the participants to collect demographic data. Future work will consider collecting demographic information to ensure a diversity of the dataset. On the modeling side, to demonstrate the validity of the database, we implemented a simple baseline for engagement estimation, which only takes the listeners' videos as input. One future direction is to incorporate the speakers' social behaviors, enabling the model to capture the dyadic interactions and entrainment between speakers and listeners. Furthermore, our current approach utilizes mean-pooling to aggregate features across the temporal dimension, potentially discarding valuable dynamic information. To address the problem, we will implement a cross-attention in our future work.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 2211550. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank our colleagues Yunweng Wang, Yuanfeixue Nan, Lynn Miller and Maja Matarić for the fruitful discussions and input. We thank Joey Cindass for contributing to developing the engagement rating scale.

REFERENCES

[1] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, 1–6.

[2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*. 354–361.

[3] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 464–472.

[4] Dan Bohus and Eric Horvitz. 2009. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference, The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 10.

[5] Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*. 2–9.

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

[7] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 350–359. https://doi.org/10.1145/3136755.3136780

[8] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[9] Ginevra Castellano, Iolanda Leite, Andre Pereira, Carlos Martinho, Ana Paiva, and Peter W McOwan. 2012. Detecting engagement in HRI: An exploration of social and task-based context. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 421–428.

[10] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2019. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. *IEEE Transactions on Affective Computing* 10, 4 (2019), 484–497. https://doi.org/10.1109/TAFFC.2017.2737019

[11] Mathieu Chollet and Stefan Scherer. 2017. Assessing Public Speaking Ability from Thin Slices of Behavior. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. 310–316. https://doi.org/10.1109/FG.2017.45

[12] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement modeling in dyadic interaction. In *2019 international conference on multimodal interaction*. 440–445.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] M. Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6, 1 (Jan. 2019), 1. https://doi.org/10.1186/s40561-018-0080-z

[15] Svati Dhamija and Terranee E Boult. 2017. Automated mood-aware engagement prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[16] Kevin Doherty and Gavin Doherty. 2018. Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.* 51, 5, Article 99 (nov 2018), 39 pages. https://doi.org/10.1145/3234149

[17] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[18] Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester. 2020. Early Prediction of Visitor Engagement in Science Museums with Multimodal Learning Analytics. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 107–116. https://doi.org/10.1145/3382507.3418890

[19] Mary Ellen Foster, Andre Gaschler, and Manuel Giuliani. 2013. How can i help you' comparing engagement classification strategies for a robot bartender. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 255–262.

[20] Alesia Gainer, Allison Aptaker, Ron Artstein, David Cobbins, Mark Core, Carla Gordon, Anton Leuski, Zongjian Li, Chirag Merchant, David Nelson, Mohammad Soleymani, and David Traum. 2023. DIVIS: Digital Interactive Victim Intake Simulator. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (<conf-loc>, <city>Würzburg</city>, <country>Germany</country>, </conf-loc>) *(IVA '23)*. Association for Computing Machinery, New York, NY, USA, Article 63, 2 pages. https://doi.org/10.1145/3570945.3607328

[21] Nadine Glas and Catherine Pelachaud. 2015. Definitions of engagement in human-agent interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 944–949. https://doi.org/10.1109/ACII.2015.7344688

[22] Joseph F Grafsgaard, Joseph B Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 42–49.

[23] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[24] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. pmlr, 448–456.

[25] Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J Matarić. 2020. Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics* 5, 39 (2020), eaaz3791.

[26] Ashish Kapoor, Rosalind W Picard, and Yuri Ivanov. 2004. Probabilistic combination of multiple modalities to detect interest. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 3. IEEE, 969–972.

[27] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. Prediction and Localization of Student Engagement in the Wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. 1–8. https://doi.org/10.1109/DICTA.2018.8615851

[28] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[29] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

[30] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.

[31] Séverin Lemaignan, Charlotte E. R. Edmunds, Emmanuel Senft, and Tony Belpaeme. 2018. The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PLOS ONE* 13, 10 (10 2018), 1–19. https://doi.org/10.1371/journal.pone.0205999

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[33] Matthew Lombard, Lisa Weinstein, and Theresa Ditton. 2011. Measuring telepresence: The validity of the Temple Presence Inventory (TPI) in a gaming context. In *ISPR 2011: The International Society for Presence Research Annual Conference*. Edinburgh UK.

[34] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. OpenReview, New Orleans, LA, USA, 18 pages.

[35] Malia F Mason, Elizabeth P Tatkow, and C Neil Macrae. 2005. The look of love: Gaze shifts and person perception. *Psychological science* 16, 3 (2005), 236–239.

[36] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2023. MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions. In *Proceedings of the 31st ACM International Conference on Multimedia* (<conf-loc>, <city>Ottawa ON</city>, <country>Canada</country>, </conf-loc>) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 9640–9645. https://doi.org/10.1145/3581783.3613851

[37] Catharine Oertel, Céline De Looze, Stefan Scherer, Andreas Windmann, Petra Wagner, and Nick Campbell. 2011. Towards the automatic detection of involvement in conversation. In *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues: COST 2102 International Conference, Budapest, Hungary, September 7-10, 2010, Revised Selected Papers*. Springer, 163–170.

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*.

[39] Karl Pearson. 1896. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 187 (1896), 253–318.

[40] Christopher Peters, Ginevra Castellano, and Sara de Freitas. 2009. An exploration of user engagement in HCI. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots* (Boston, Massachusetts) *(AFFINE '09)*. Association for Computing Machinery, New York, NY, USA, Article 9, 3 pages. https://doi.org/10.1145/1655260.1655269

[41] Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision.

In *International Conference on Machine Learning*. PMLR, 28492–28518.

[43] Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances* 9, 13 (2023), eadf3197.

[44] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L. Sidner. 2010. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 375–382. https://doi.org/10.1109/HRI.2010.5453163

[45] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–8. https://doi.org/10.1109/FG.2013.6553805

[46] Hanan Salam, Oya Celiktutan, Hatice Gunes, and Mohamed Chetouani. 2023. Automatic Context-Aware Inference of Engagement in HMI: A Survey. *IEEE Transactions on Affective Computing* (2023), 1–20. https://doi.org/10.1109/TAFFC.2023.3278707

[47] Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2016. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access* 5 (2016), 705–721.

[48] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 252, 13 pages. https://doi.org/10.1145/3411764.3445615

[49] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne, Switzerland) *(HRI '11)*. Association for Computing Machinery, New York, NY, USA, 305–312. https://doi.org/10.1145/1957656.1957781

[50] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*. 305–312.

[51] Monisha Singh, Ximi Hoque, Donghuo Zeng, Yanan Wang, Kazushi Ikeda, and Abhinav Dhall. 2023. Do I Have Your Attention: A Large Scale Engagement Prediction Dataset and Baselines. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 174–182.

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[53] Baltrusaitis Tadas, Zadeh Amir, Lim Yao Chong, and Louis-Morency Philippe. 2018. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*.

[54] Minh Tran, Yufeng Yin, and Mohammad Soleymani. 2023. Personalized Adaptation with Pre-trained Speech Encoders for Continuous Emotion Recognition. *arXiv preprint arXiv:2309.02418* (2023).

[55] 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.

[56] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE international conference on computer vision*. 3756–3764.

[57] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051* (2021).

[58] Yufeng Yin, Jiashu Xu, Tianxin Zu, and Mohammad Soleymani. 2022. X-Norm: Exchanging Normalization Parameters for Bimodal Fusion. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 605–614.

[59] Chen Yu, Paul M Aoki, and Allison Woodruff. 2004. Detecting user engagement in everyday conversations. *arXiv preprint cs/0410027* (2004).

[60] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2519–2528.