Tracking Engagement in Real-world Human Communication from Wearable Sensors

Sunreeta Bhattacharya[1,4], Kevin Joo[2], Cheng Ma[2], Alvaro Fernandez[2], Yiming Fang[2], Fernando De la Torre[2], Lori L. Holt[3]

[1]Neuroscience Institute, Carnegie Mellon University, [2]Robotics Institute, Carnegie Mellon University, [3]Dept of Psychology, University of Texas, Austin, [4]Center for the Neural Basis of Cognition, Pittsburgh

Introduction

- Most studies of human communication take place in controlled environments rather than in the wild
- Using novel wearable sensor technology, we devise a new way of studying naturalistic communication in human dyads
- We target dyadic engagement to identify relevant paralinguistic features in speech that are related to the outcome of an interaction

Study Design and Apparatus

Pupil Invisible™ glasses

- Participants wore **eye-tracking** glasses with **egocentric video cameras** and **microphone**.

Side-on view of a dyad

POV from each egocentric camera

Design:

- Two strangers in a spacious setting outside the lab
- 10–15-minute conversation
- Shared personal experiences during the lockdowns, exchanged views on vaccine mandates

- *"Did you establish a good rapport with your partner?"*
- *"Was there a natural back-and-forth in the conversation?"*
- ...

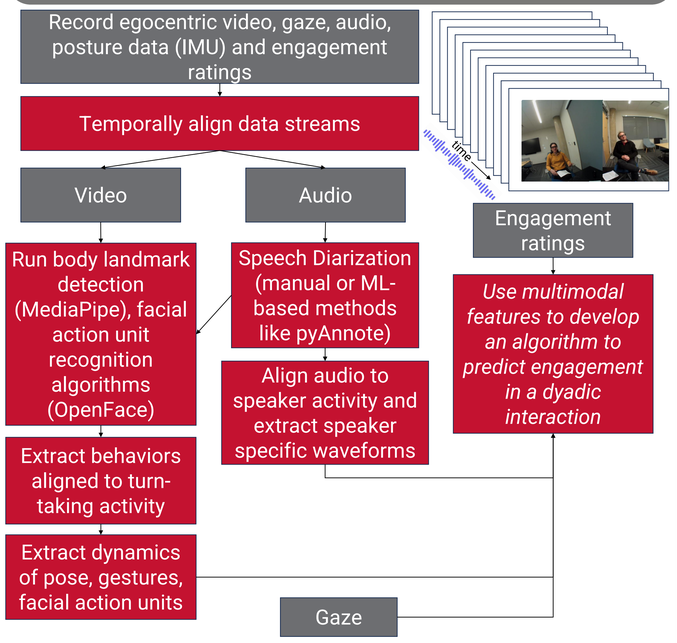
Sample questions from the engagement questionnaire

We recorded:
- Speech
- Egocentric video
- Gaze
- Movement (IMU data)

with informed consent of n=19 dyads.

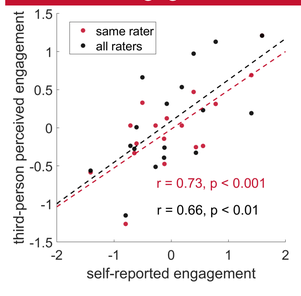
- 53-item questionnaire captured level of engagement during the interaction.

Data Processing



Results

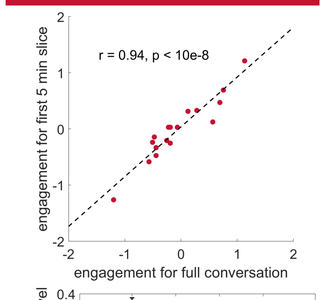
External raters of engagement agree with first-person reports of engagement

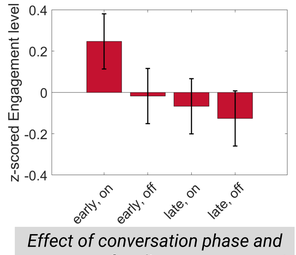


The speech signal affects engagement levels of an external observer
Earlier parts of the conversation appear to be more engaging than later parts

Third-person raters watched context-free conversation excerpts (10-sec/1-min) **with** or **without** audio and reported dyadic engagement.

Engagement scores for first impressions are correlated with full-session scores

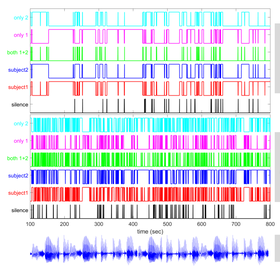




Effect of conversation phase and presence of audio on engagement (error bars indicate SEM)

Results - continued

Speech diarization assigns speaker identity to the speech waveform



*Turn-taking profile in **low engagement** session*

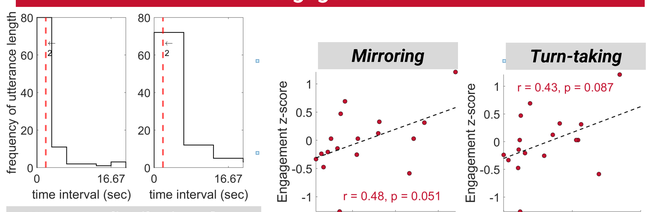
*Turn-taking profile in **high engagement** session*

Speech

- Automatic diarization was performed using pyAnnote.
- Outputs turn timestamps with speaker information.
- Overlaps and silent periods can be extracted using logic operations.

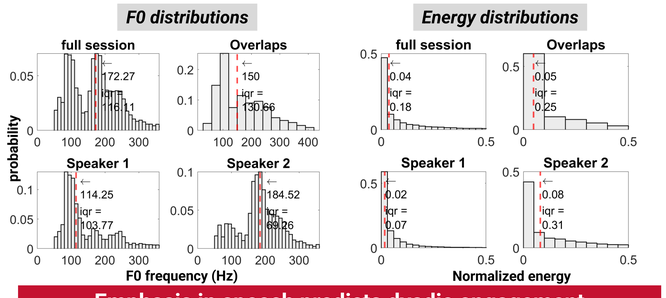
Visualization of turn-taking data obtained through speaker diarization. (above) This is used to align the speech waveform (below).

Mirroring and turn-taking behavior predict dyadic engagement

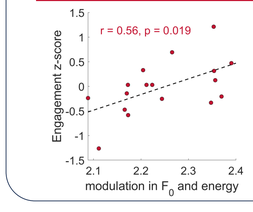


Utterance distributions from turn-taking data. Short utterances less than 2 seconds long were counted as mirroring events considered to be a means of backchannel communication.

Mirroring behavior and turn-taking occur more frequently in engaging conversations.

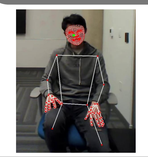


Emphasis in speech predicts dyadic engagement



Emphasis = sum of variability observed in each participant's tone (F0) and intensity (energy). This modulation was quantified as the sum of the variability in median-normalized F0 and energy.

Ongoing and Future Work



Facemesh and body landmarks (using MediaPipe) in red and gaze (in green) → *identify FAUs (facial action units) and gestures*

- What visual features (such as facial expressions, gestures) are being attended to – as tracked by gaze location?
- Can we train an ML model to detect engaging parts of a conversation?
- Can language models predict and reason about engagement in dyads?

Summary

- Found robust markers of communicative engagement in non-verbal features of speech:
  - voice modulations - the variance of speaker-separated pitch and energy distributions
  - extent of coupling measured as frequency of acoustic mirroring events
  - the frequency of turn-taking events.
- Replicated classic work on naturalistic speech in dyads (e.g. using hand coding of features via the 'sociometer'[2]) in naturalistic setups
- Developed a pipeline useful for human-robot interactions

References

1. Curhan, J. R. & Pentland, A. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *J Appl Psychol* **92**, 802–811 (2007).
2. Duncan, S. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* **23**, 283–292 (1972).
3. pyAnnote, Herve Bredin et al, 2019

Acknowledgements

We are grateful to all those who helped us on this project, especially Erin Smith and Christi Gomez of the Holt Lab, and Kailana Baker-Matsuoka for her work on video ratings.

This work was supported by a James S. McDonnell Foundation grant to LLH and FDlT.