# Multimodal Sequential Modeling of Task-Mediated Frustration

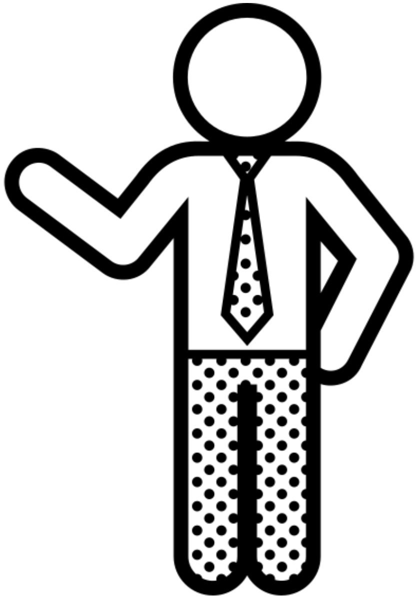
*Intermediate Fusion for High and Low Level Features*

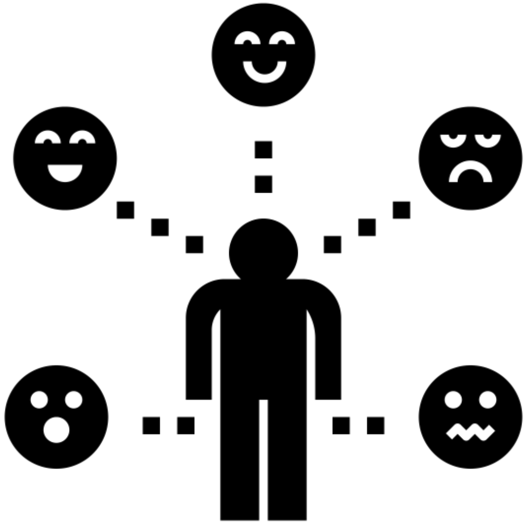
**Michael Peechatt**

ENGL.584.01/684.01 - Speech Processing II

CLaSP Lab + Graphics Lab @ RIT

# What is Multimodality?





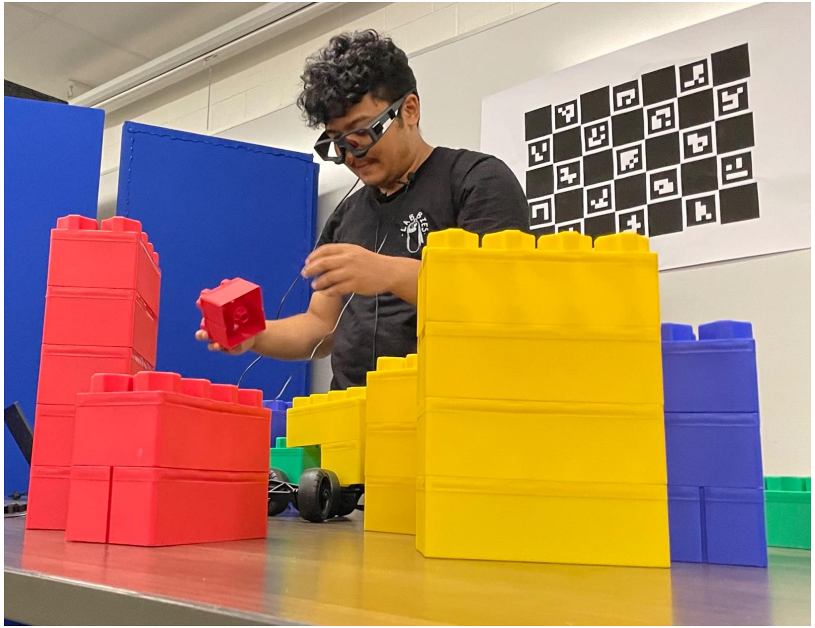
# Motivation

- ❑ **Cooperation is under-explored in affective computing**
  - ❑ The pandemic increased Zoom usage in an collaborative context
- ❑ **Human perception combines low and high level features**
  - ❑ Can our machine learning models reflect this?
- ❑ **RQ1:** *Is or high or low level features better for identifying frustration?*
- ❑ **RQ2:** *Does, on average, fusing predictions improve overall performance?*

# Setup

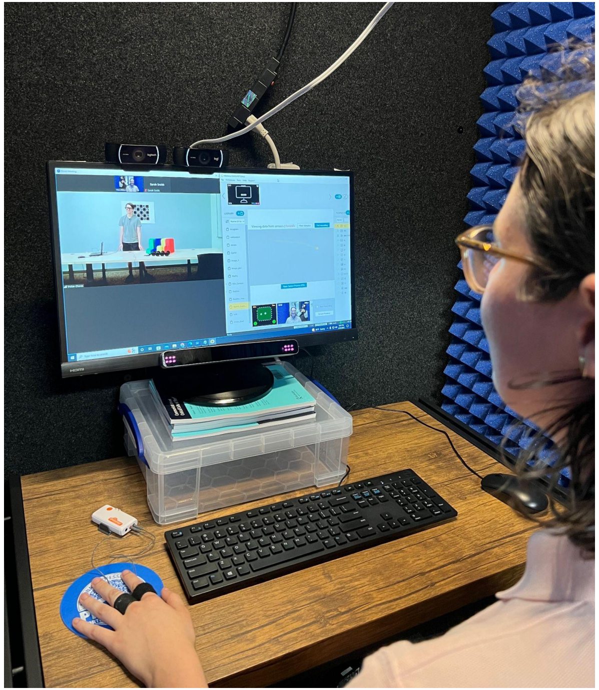
## Builder

*in CLaSP Lab*

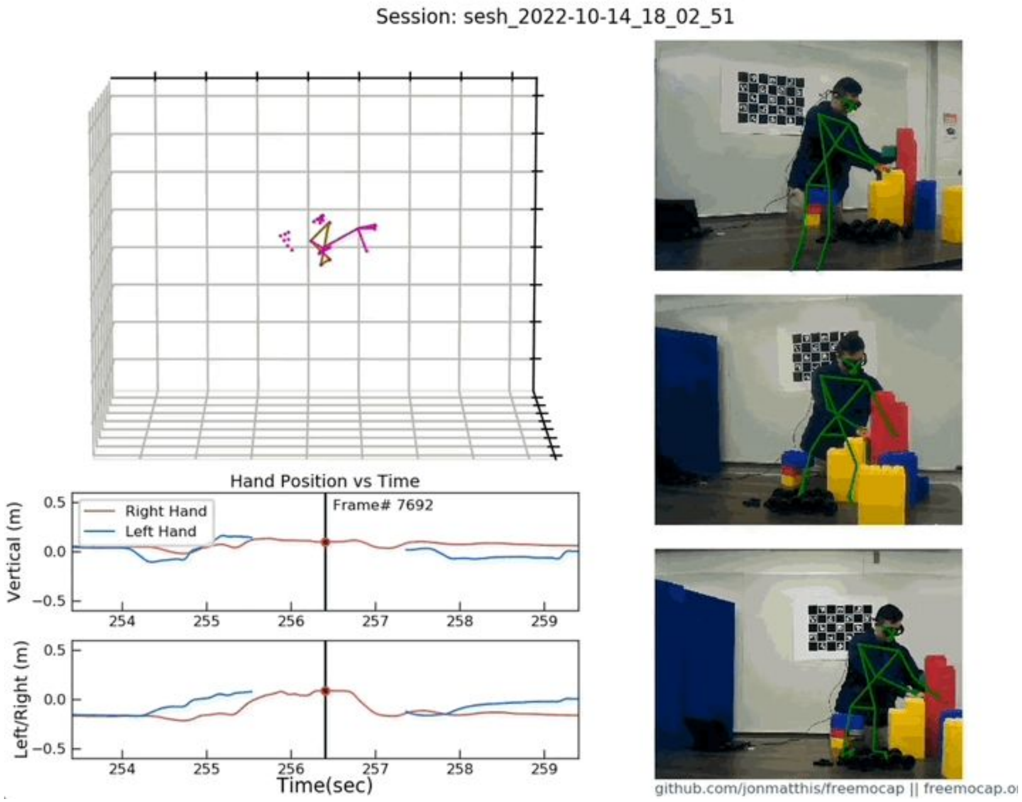


## Instructor

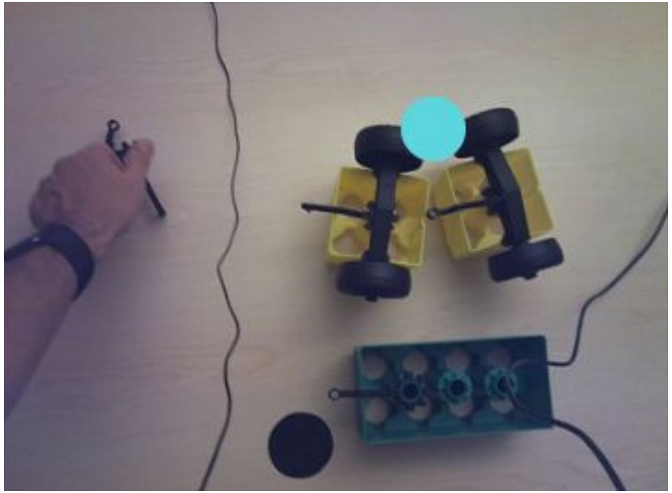
*in Whisper Room*



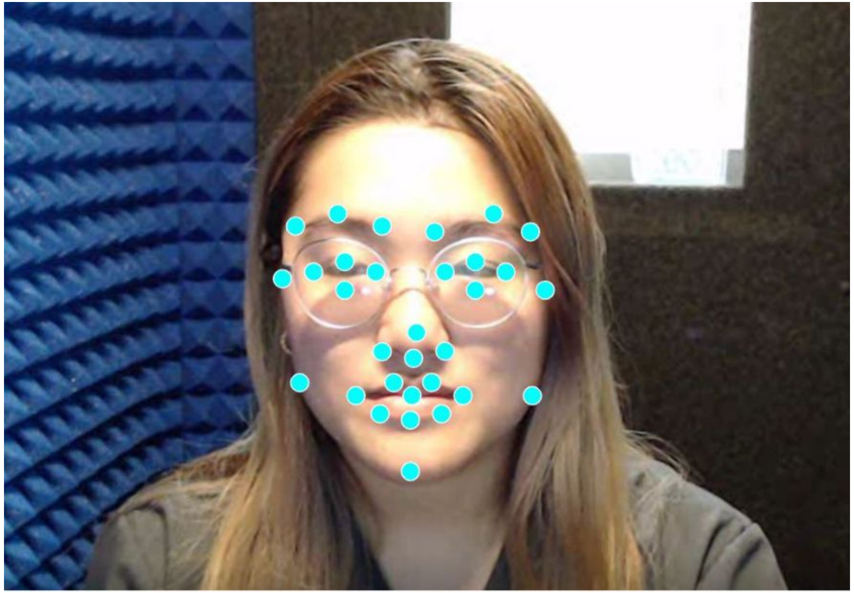
# Builder Modalities



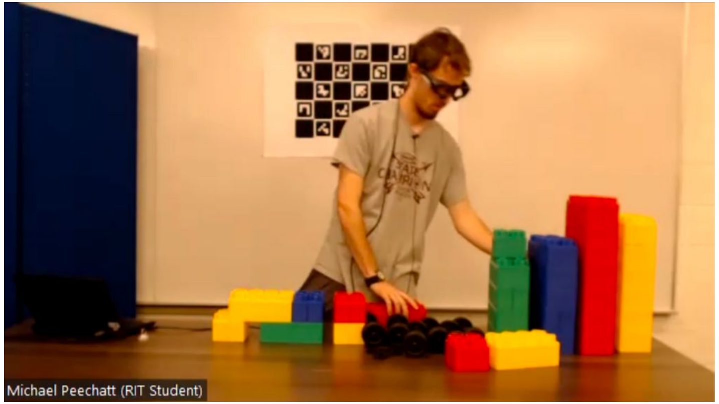
github.com/jonmatthis/freemocap || freemocap.o

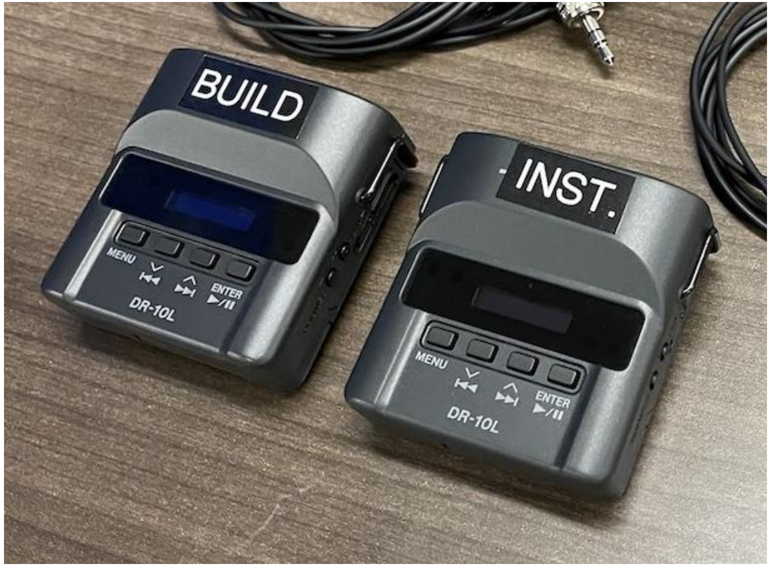


# Instructor Modalities

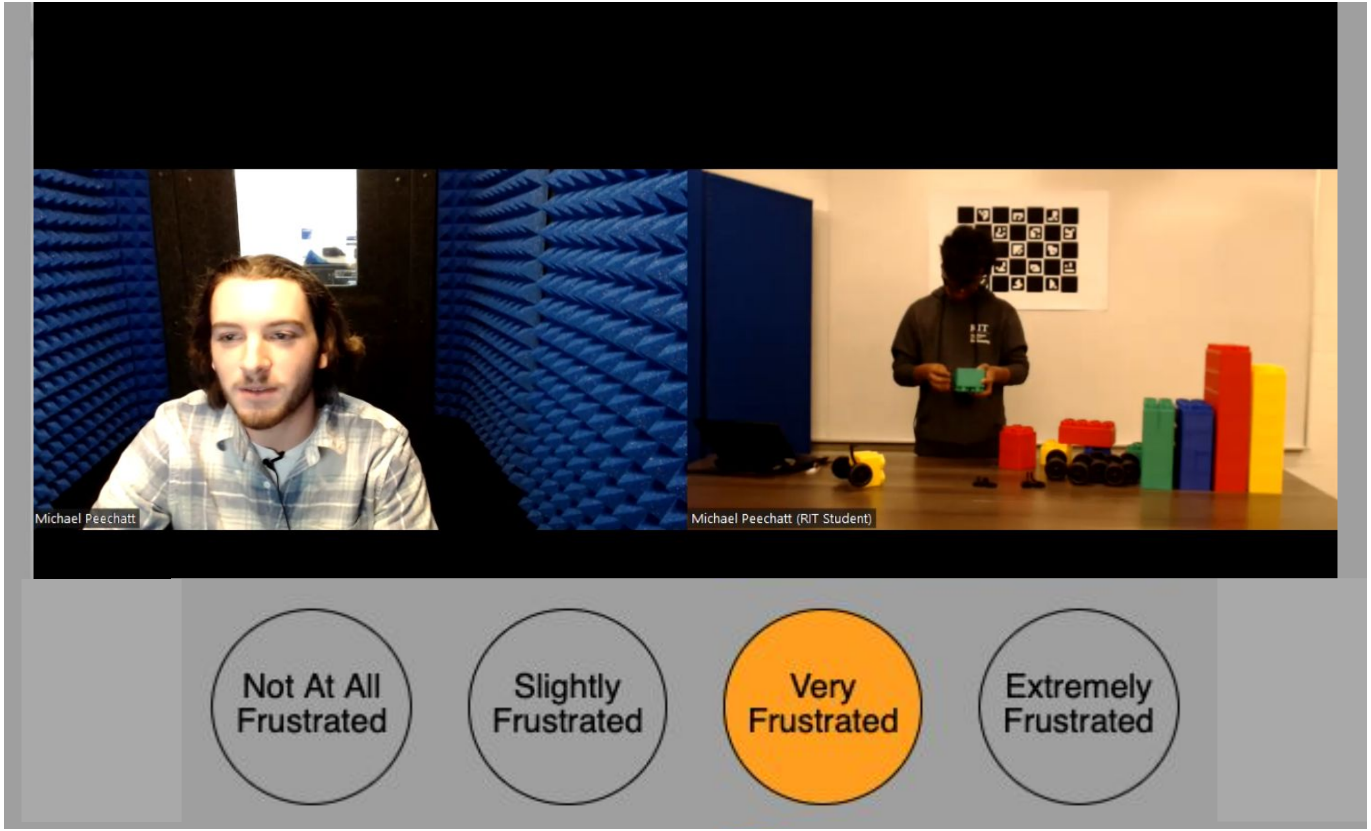


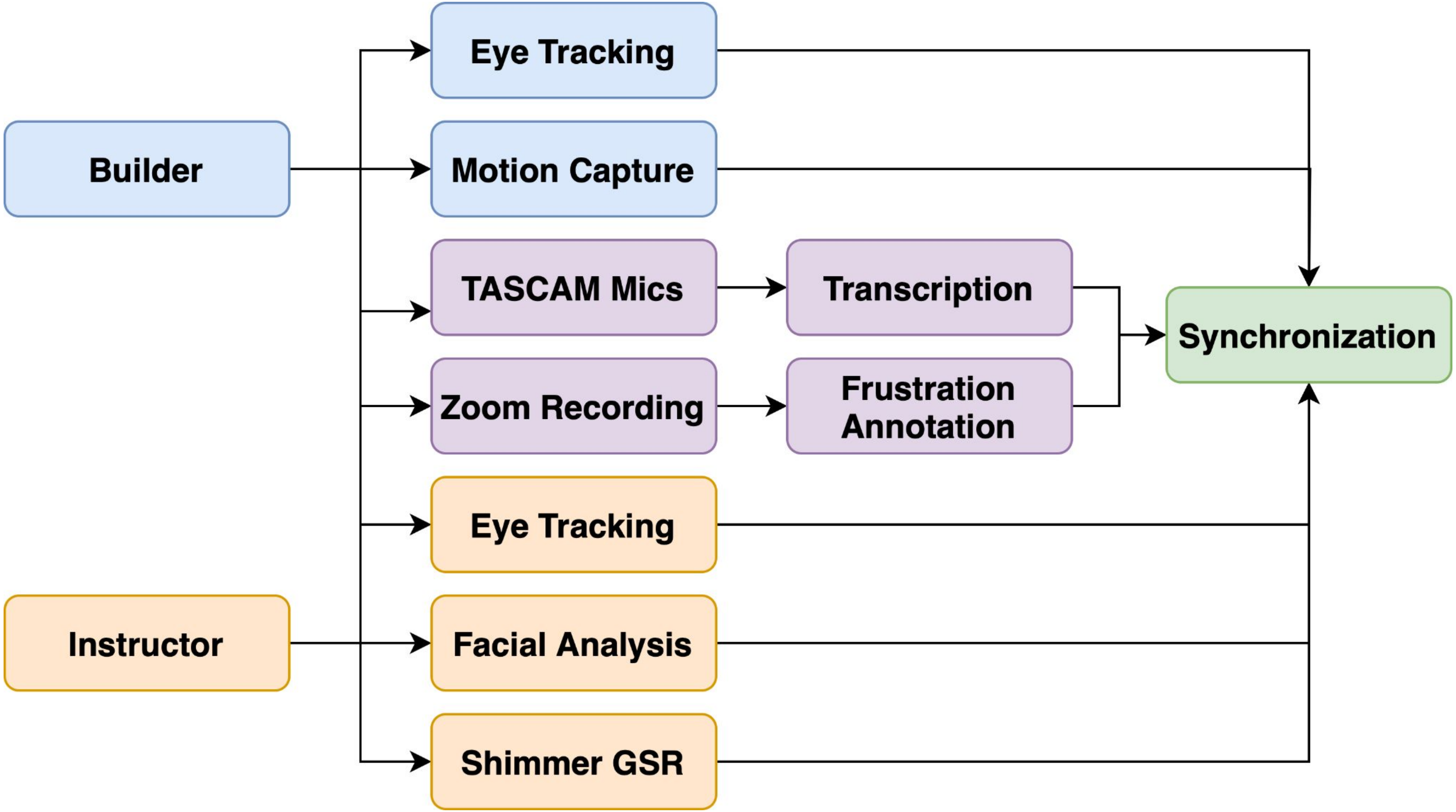
# Common Modalities





# Annotation

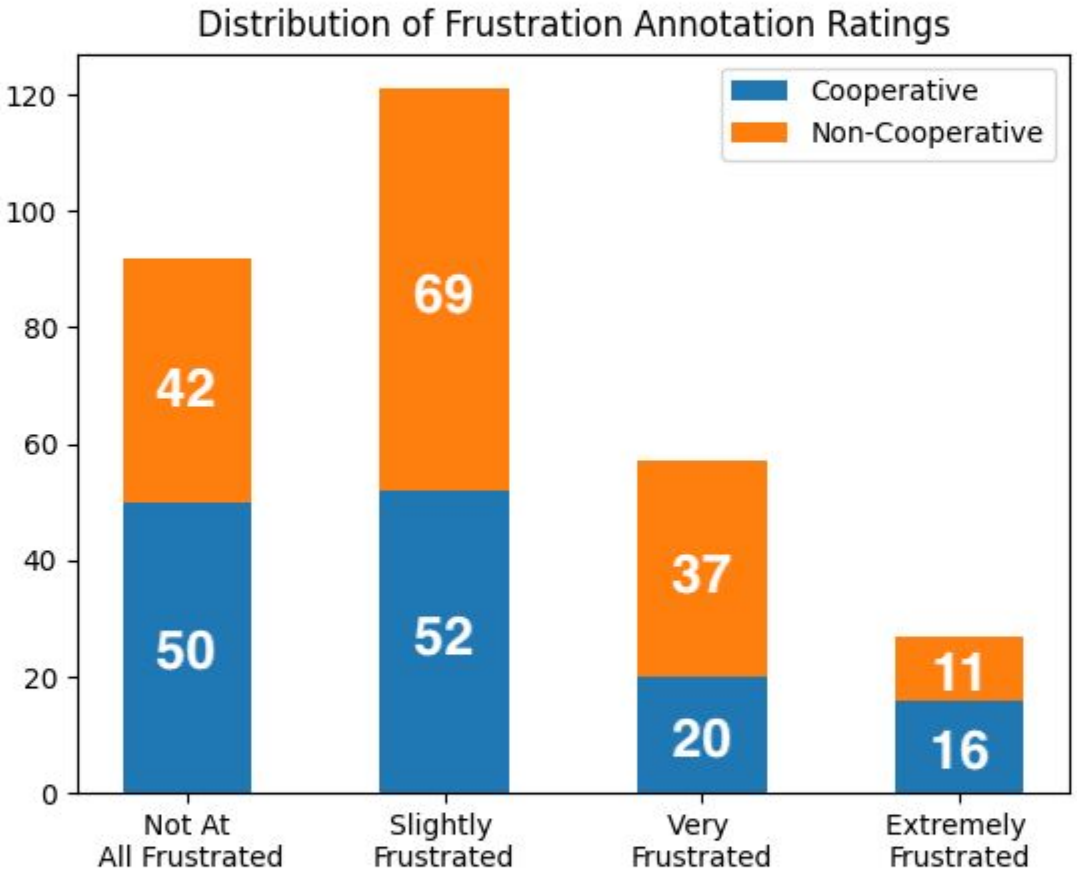

Michael Peechatt
Michael Peechatt (RIT Student)
Not At All Frustrated
Slightly Frustrated
Very Frustrated
Extremely Frustrated

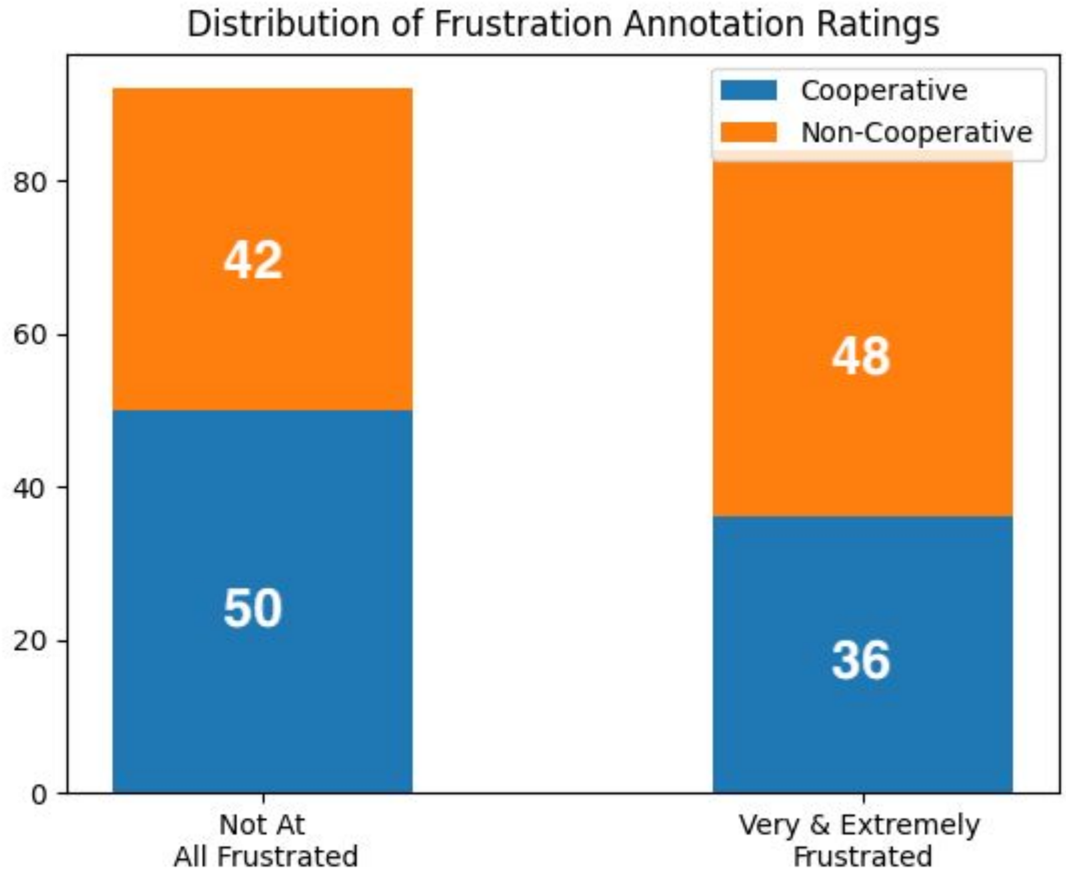

Eye Tracking
Builder
Motion Capture
TASCAM Mics
Transcription
Synchronization
Zoom Recording
Frustration Annotation
Eye Tracking
Instructor
Facial Analysis
Shimmer GSR

# MULTICOLLAB Dataset

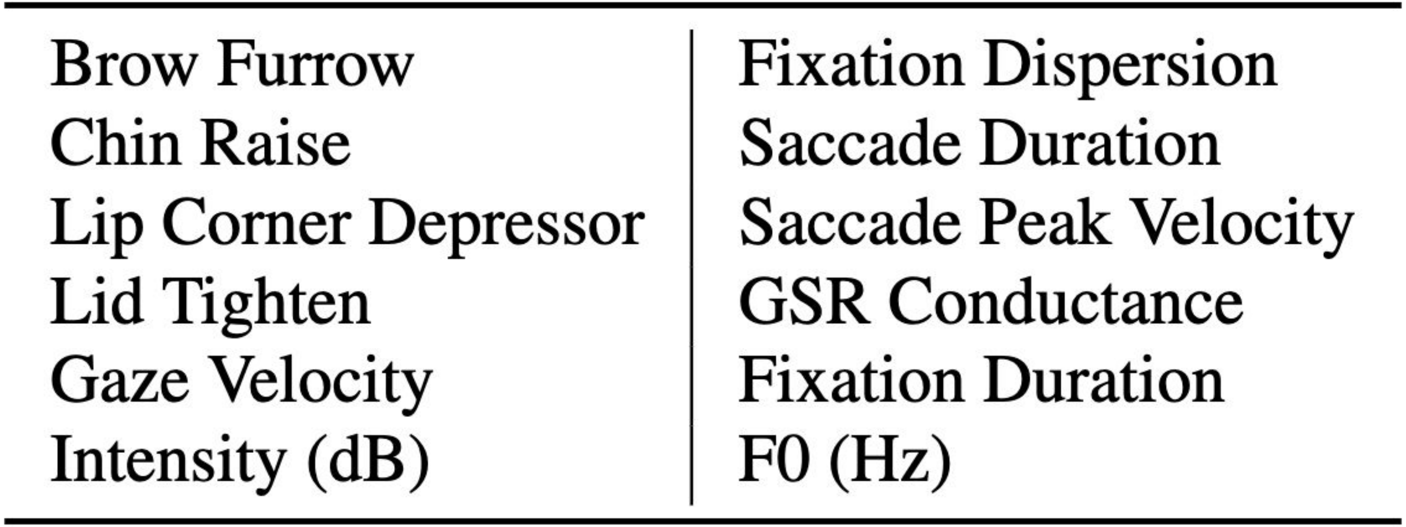
- ❑ **48 subjects (24 builder-instructor groups)**
  - ❑ 42% Female, 56% Male, 2% undisclosed
  - ❑ 8 groups had different gendered interactions
  - ❑ 16 groups had same gendered interactions

- ❑ **Ethnicity Distribution**
  - ❑ 2.1% Southeast Asian, 8.4% African-American, 10.5% Hispanic, 39.6% Asian, and 37.5% Caucasian (1.8% Undisclosed)
  - ❑ 20.8% ESL speakers, 79.2% native English speakers
  - ❑ 3 of 24 groups were mix of non-native and native interactions


Distribution of Frustration Annotation Ratings
Cooperative
Non-Cooperative
0
20
40
60
80
100
120
50
42
52
69
20
37
16
11
Not At All Frustrated
Slightly Frustrated
Very Frustrated
Extremely Frustrated

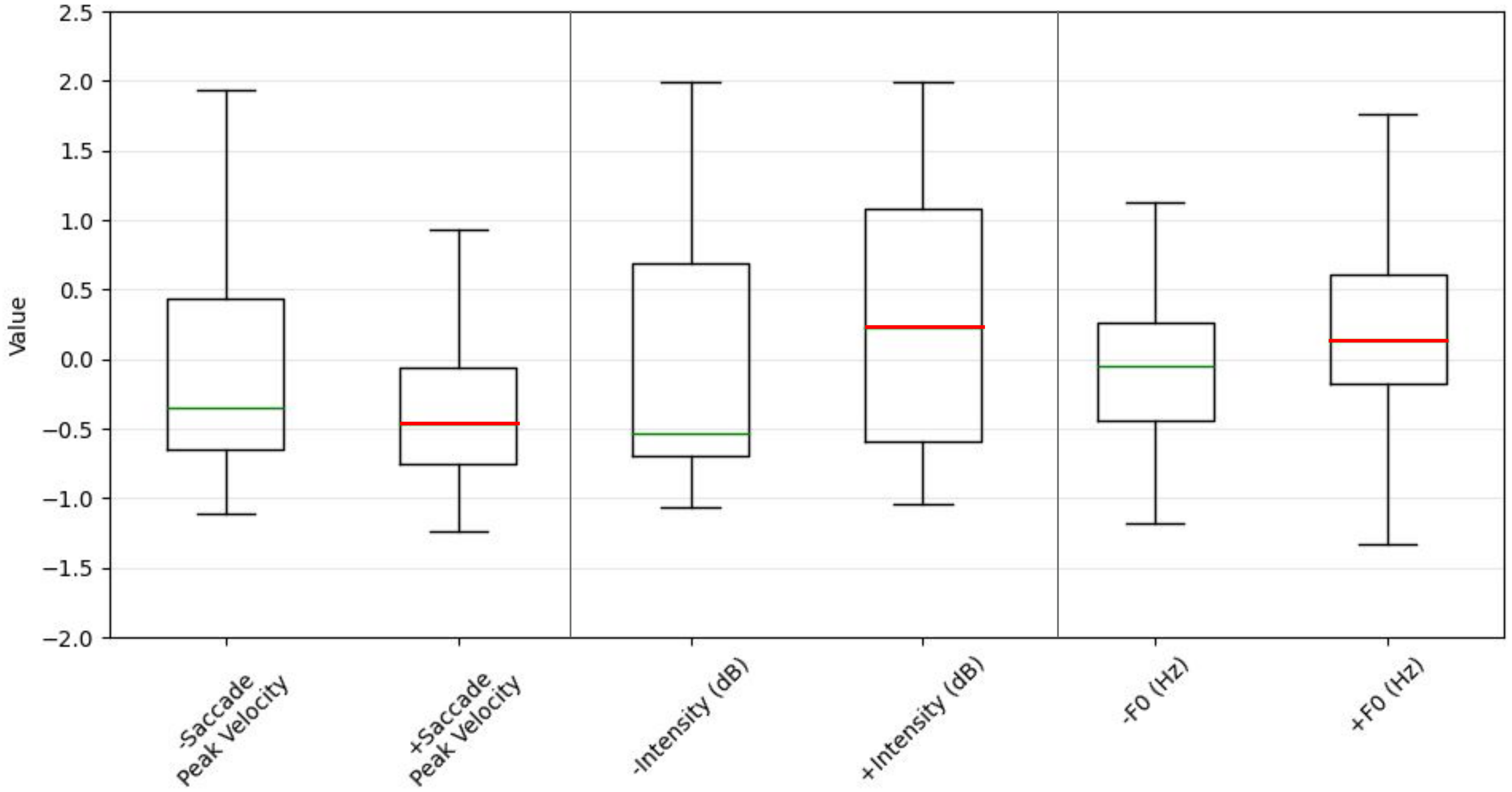

Distribution of Frustration Annotation Ratings
Cooperative
Non-Cooperative
0
20
40
60
80
42
50
48
36
Not At
All Frustrated
Very & Extremely
Frustrated

# Feature Extraction



| | |
|---|---|
| Brow Furrow | Fixation Dispersion |
| Chin Raise | Saccade Duration |
| Lip Corner Depressor | Saccade Peak Velocity |
| Lid Tighten | GSR Conductance |
| Gaze Velocity | Fixation Duration |
| Intensity (dB) | F0 (Hz) |

Z-score normalized


Value
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
-Saccade Peak Velocity
+Saccade Peak Velocity
-Intensity (dB)
+Intensity (dB)
-F0 (Hz)
+F0 (Hz)

# Dataset Shape

$$\begin{array}{cc}
m_1 \qquad m_2 \qquad \cdots \qquad m_n & \\
\left[\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_n \end{array}\right] & \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_n \end{array}
\end{array}$$

# Dataset Shape

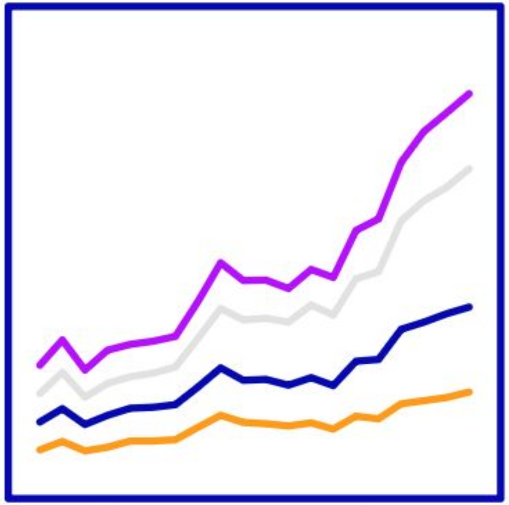
$$\begin{array}{c} \begin{array}{cccc} m_1 & m_2 & \cdots & m_n \end{array} \\ \left[\begin{array}{cccc} \quad & \quad & t_1 & \quad \\ & & t_2 & \\ & & \vdots & \\ & & t_n & \end{array}\right] \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_n \end{array} \end{array}$$

$$\Big\downarrow \text{flatten}$$

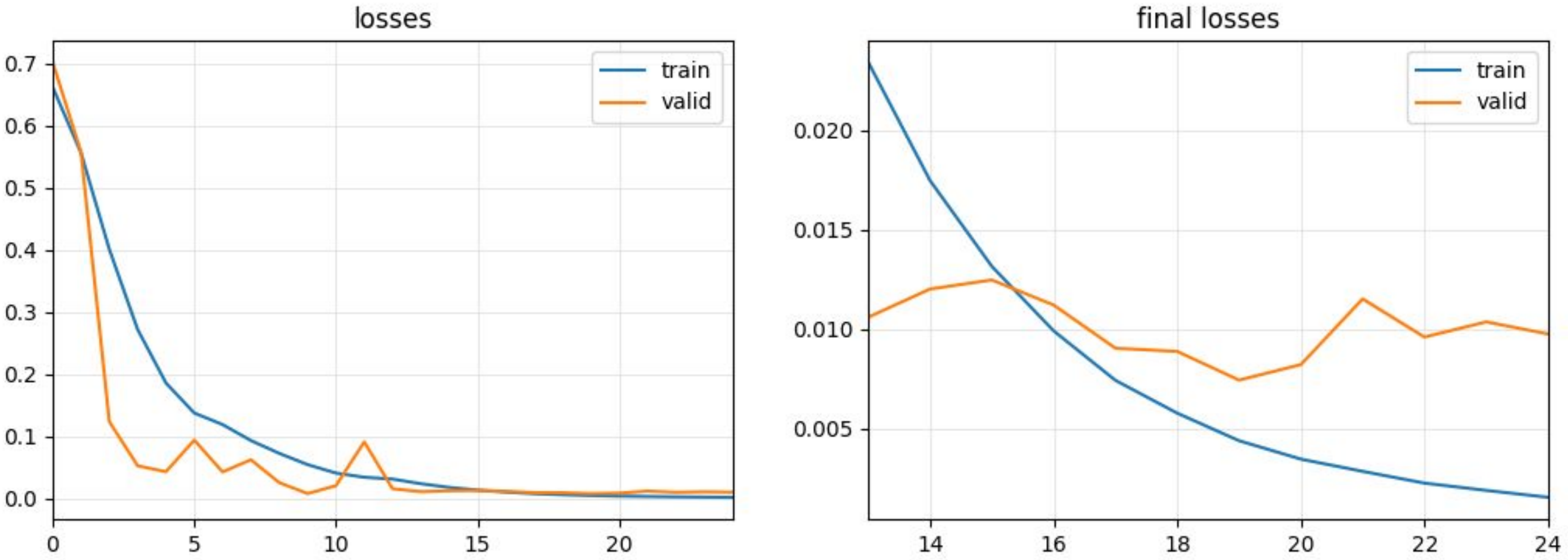
$$||\{m_1, m_2, \ldots, m_n\}|| \cdot t_n$$

$$\begin{array}{c} \begin{array}{cccc} m_1\_t_1 & m_2\_t_1 & \cdots & m_n\_t_n \end{array} \\ \left[\begin{array}{cccc} \quad & \quad & \quad & \quad \\ & & & \end{array}\right] \end{array}$$

# tsai



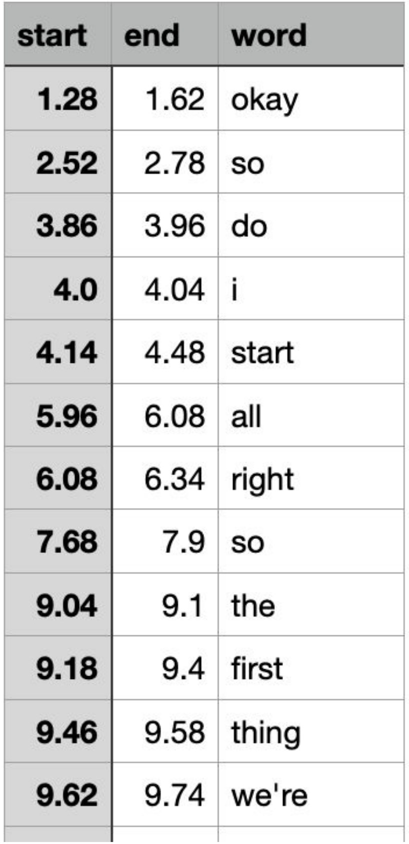
state-of-the-art deep learning
for time series and sequences



t_win = 10000 milliseconds, t_n = 20

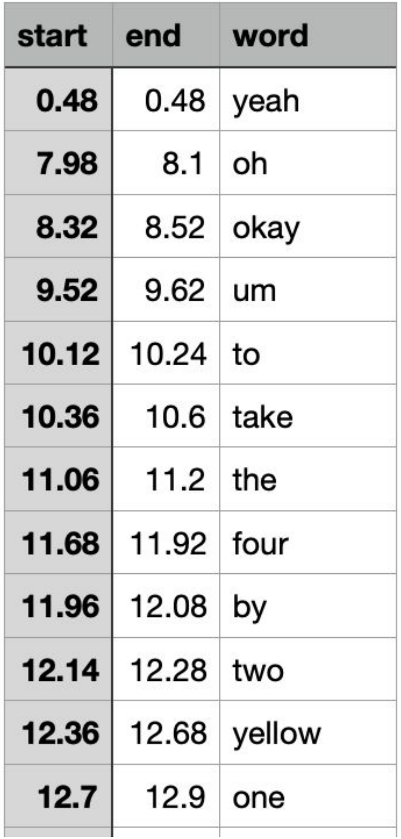
avg. accuracy = 0.523

30NI_transcript



| start | end | word |
|---|---|---|
| 1.28 | 1.62 | okay |
| 2.52 | 2.78 | so |
| 3.86 | 3.96 | do |
| 4.0 | 4.04 | i |
| 4.14 | 4.48 | start |
| 5.96 | 6.08 | all |
| 6.08 | 6.34 | right |
| 7.68 | 7.9 | so |
| 9.04 | 9.1 | the |
| 9.18 | 9.4 | first |
| 9.46 | 9.58 | thing |
| 9.62 | 9.74 | we're |

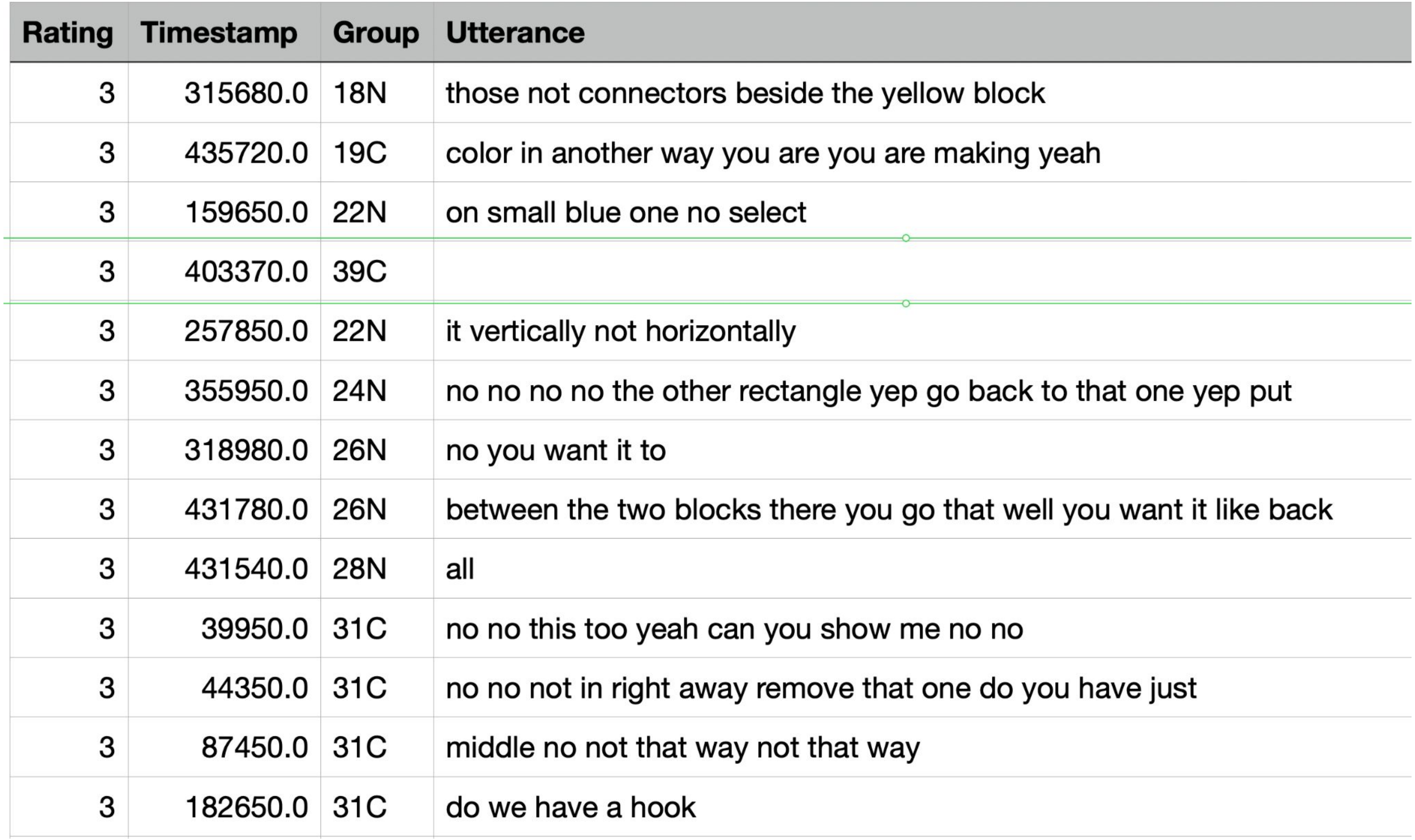
27CI_transcript



| start | end | word |
|---|---|---|
| 0.48 | 0.48 | yeah |
| 7.98 | 8.1 | oh |
| 8.32 | 8.52 | okay |
| 9.52 | 9.62 | um |
| 10.12 | 10.24 | to |
| 10.36 | 10.6 | take |
| 11.06 | 11.2 | the |
| 11.68 | 11.92 | four |
| 11.96 | 12.08 | by |
| 12.14 | 12.28 | two |
| 12.36 | 12.68 | yellow |
| 12.7 | 12.9 | one |

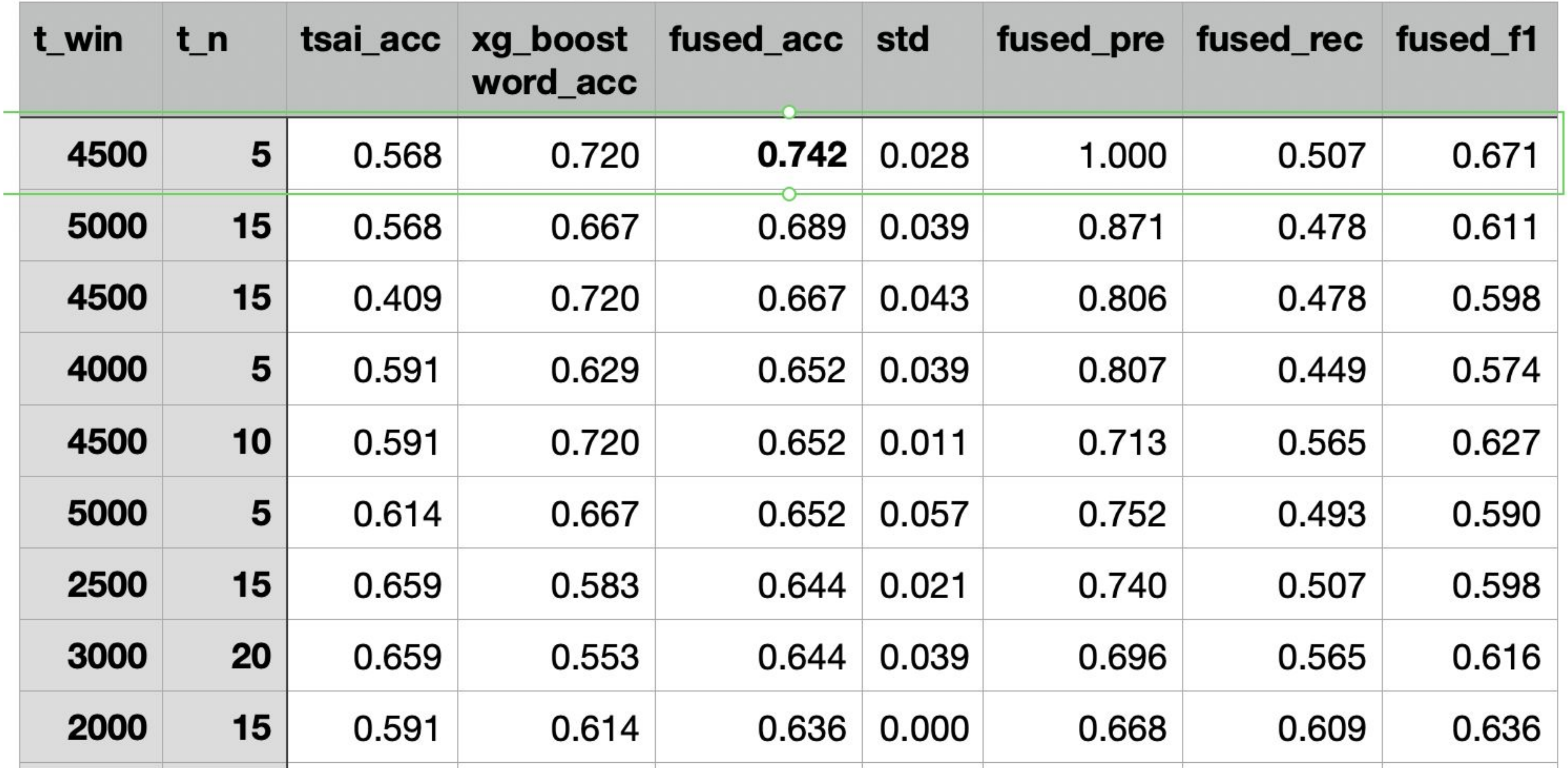
fastText

$$\frac{\sum_{w \in W} f_t(w)}{|W|}$$



| Rating | Timestamp | Group | Utterance |
|---|---|---|---|
| 3 | 315680.0 | 18N | those not connectors beside the yellow block |
| 3 | 435720.0 | 19C | color in another way you are you are making yeah |
| 3 | 159650.0 | 22N | on small blue one no select |
| 3 | 403370.0 | 39C | |
| 3 | 257850.0 | 22N | it vertically not horizontally |
| 3 | 355950.0 | 24N | no no no no the other rectangle yep go back to that one yep put |
| 3 | 318980.0 | 26N | no you want it to |
| 3 | 431780.0 | 26N | between the two blocks there you go that well you want it like back |
| 3 | 431540.0 | 28N | all |
| 3 | 39950.0 | 31C | no no this too yeah can you show me no no |
| 3 | 44350.0 | 31C | no no not in right away remove that one do you have just |
| 3 | 87450.0 | 31C | middle no not that way not that way |
| 3 | 182650.0 | 31C | do we have a hook |

# Averaging TSAI Inferences with XGBoost Word Inferences



| t_win | t_n | tsai_acc | xg_boost word_acc | fused_acc | std | fused_pre | fused_rec | fused_f1 |
|---|---|---|---|---|---|---|---|---|
| 4500 | 5 | 0.568 | 0.720 | **0.742** | 0.028 | 1.000 | 0.507 | 0.671 |
| 5000 | 15 | 0.568 | 0.667 | 0.689 | 0.039 | 0.871 | 0.478 | 0.611 |
| 4500 | 15 | 0.409 | 0.720 | 0.667 | 0.043 | 0.806 | 0.478 | 0.598 |
| 4000 | 5 | 0.591 | 0.629 | 0.652 | 0.039 | 0.807 | 0.449 | 0.574 |
| 4500 | 10 | 0.591 | 0.720 | 0.652 | 0.011 | 0.713 | 0.565 | 0.627 |
| 5000 | 5 | 0.614 | 0.667 | 0.652 | 0.057 | 0.752 | 0.493 | 0.590 |
| 2500 | 15 | 0.659 | 0.583 | 0.644 | 0.021 | 0.740 | 0.507 | 0.598 |
| 3000 | 20 | 0.659 | 0.553 | 0.644 | 0.039 | 0.696 | 0.565 | 0.616 |
| 2000 | 15 | 0.591 | 0.614 | 0.636 | 0.000 | 0.668 | 0.609 | 0.636 |

# Research Question Answers

- ❑ **Low Level Average**
  - ❑ TSAI Accuracy = 0.545
- ❑ **High Level Average**
  - ❑ XGBoost Accuracy = **0.609**
- ❑ **Fusion Average**
  - ❑ TSAI + XGBoost Accuracy = 0.595

# Future Work

- ❑ **Perform ablation study on audio features**
  - ❑ Consider into including jitter and shimmer features
- ❑ **Explore other word embeddings**
  - ❑ FastText is not optimized for *spoken dialogue*
- ❑ **Look into clustering data for generating labels**
  - ❑ Rather than relying on human annotation

# Questions?