Multimodal Sequential Modeling of Task-Mediated Frustration

Intermediate Fusion for High and Low Level Features

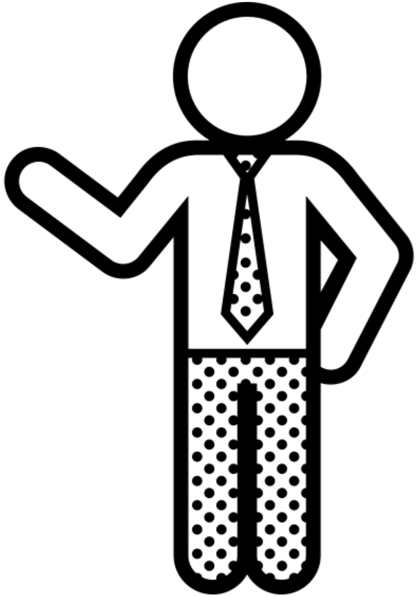
Michael Peechatt

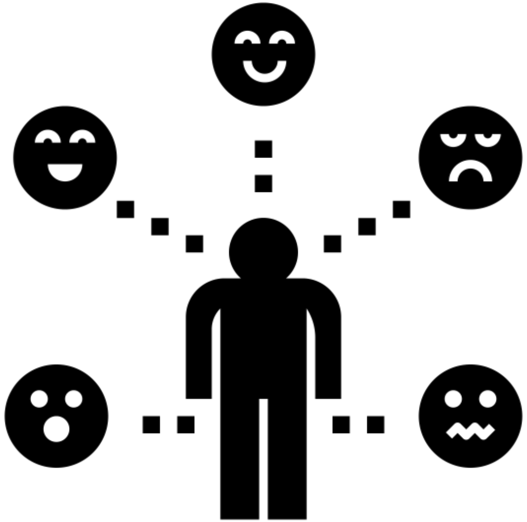
ENGL.584.01/684.01 - Speech Processing II

CLaSP Lab + Graphics Lab @ **RIT**

RIT

What is Multimodality?





Motivation

- ❑ **Cooperation is under-explored in affective computing**
  - ❑ The pandemic increased Zoom usage in an collaborative context

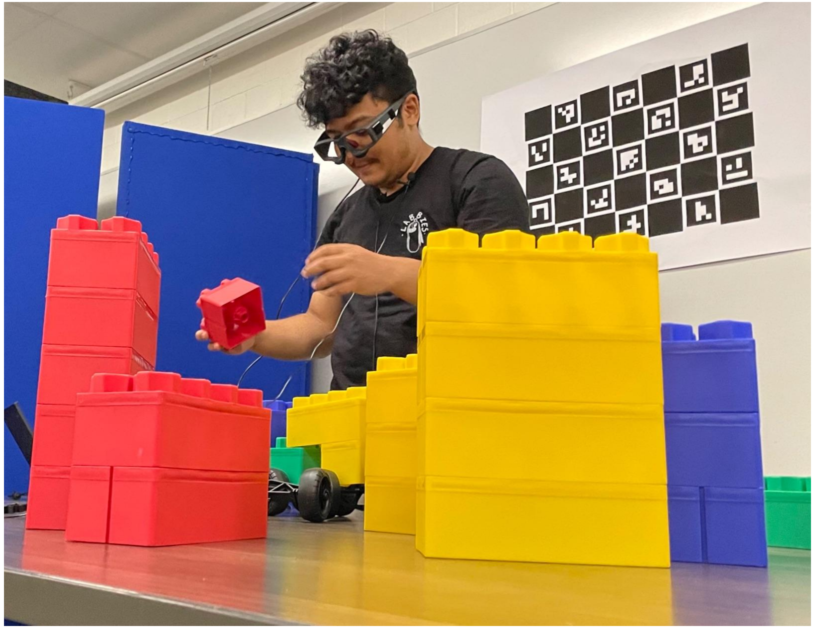
- ❑ **Human perception combines low and high level features**
  - ❑ Can our machine learning models reflect this?

- ❑ **RQ1:** *Is or high or low level features better for identifying frustration?*
- ❑ **RQ2:** *Does, on average, fusing predictions improve overall performance?*

Setup

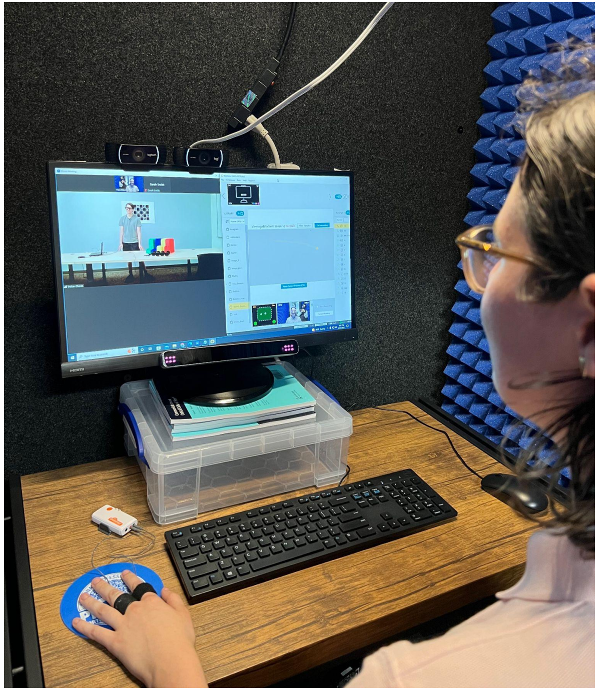
Builder

in CLaSP Lab

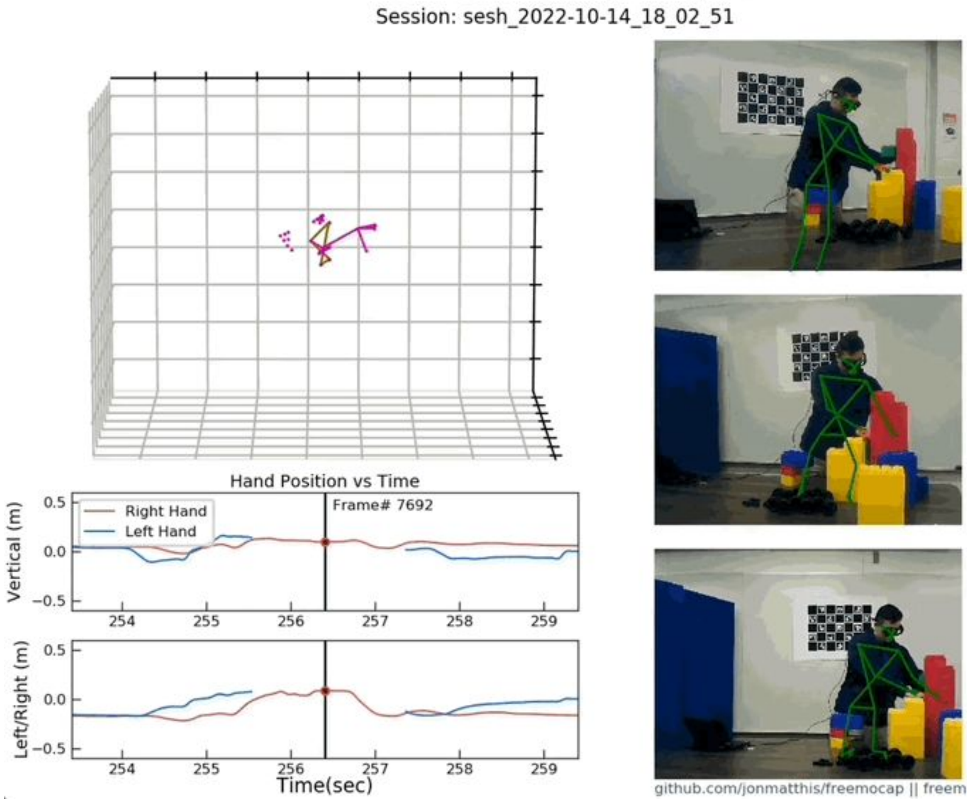


Instructor

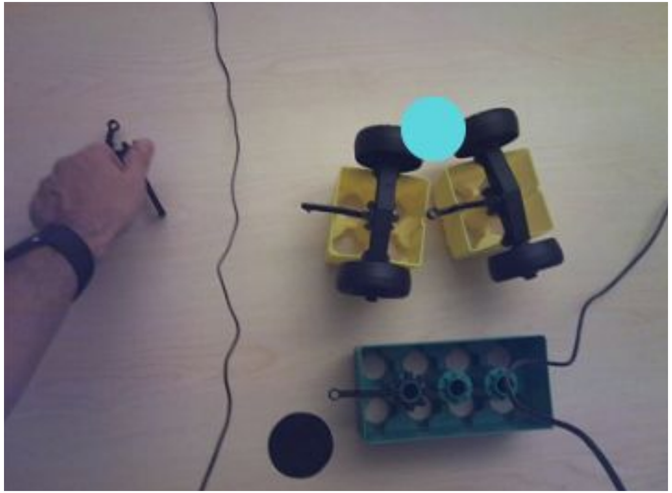
in Whisper Room



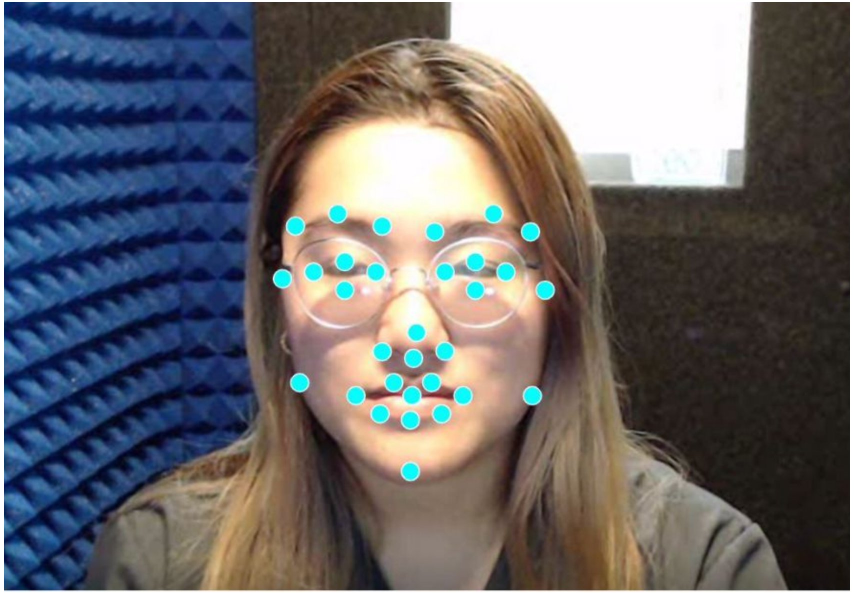
Builder Modalities



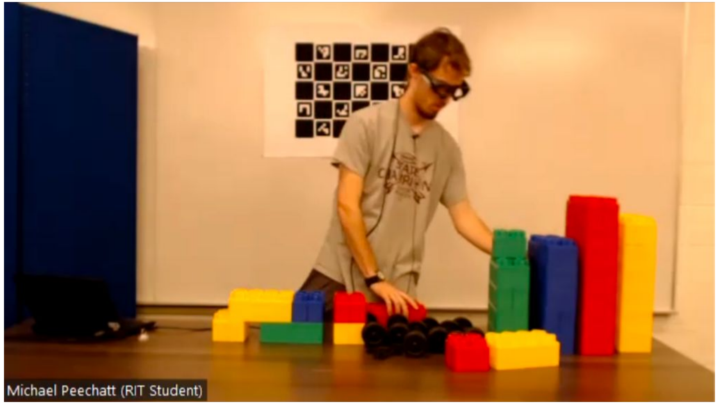
github.com/jonmatthis/freemocap || freemocap.o



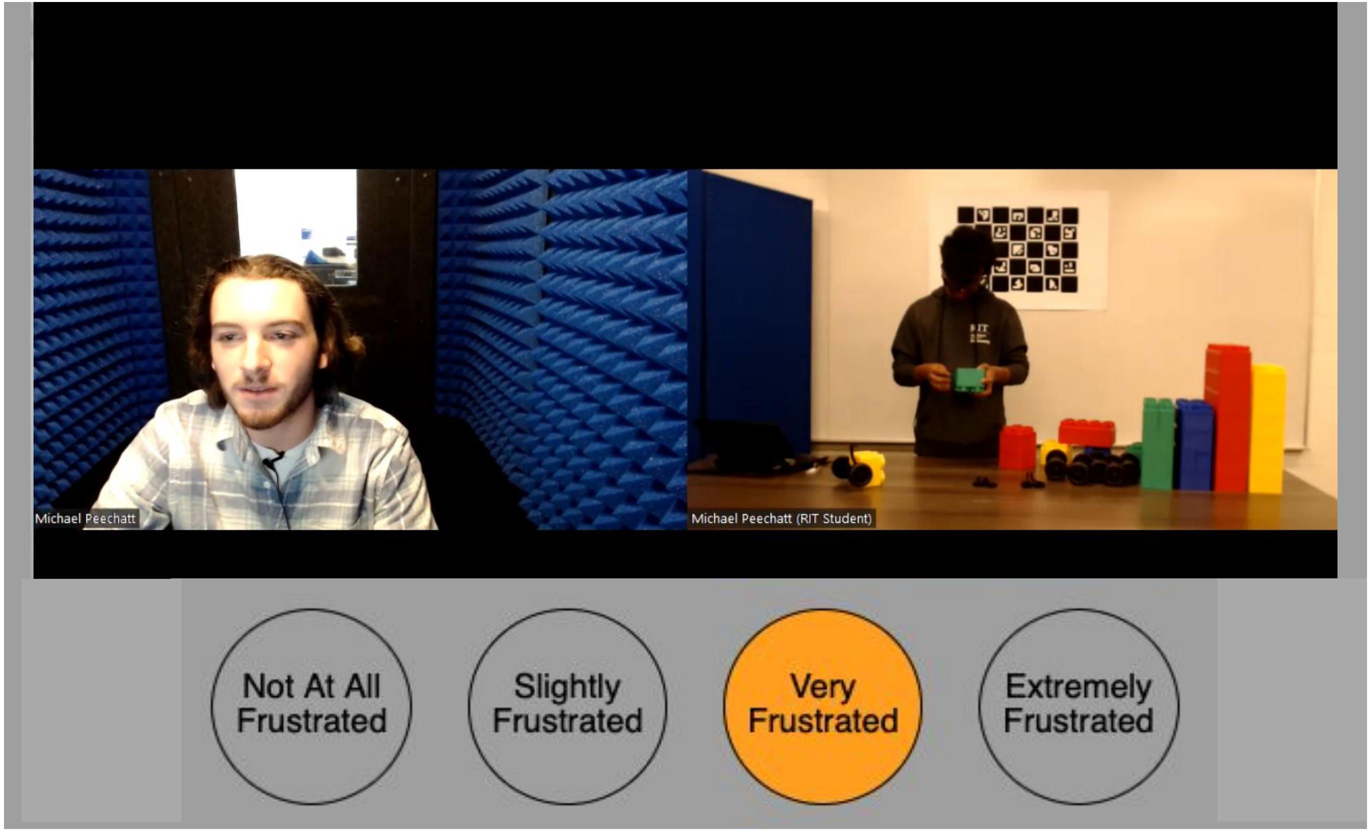
Instructor Modalities

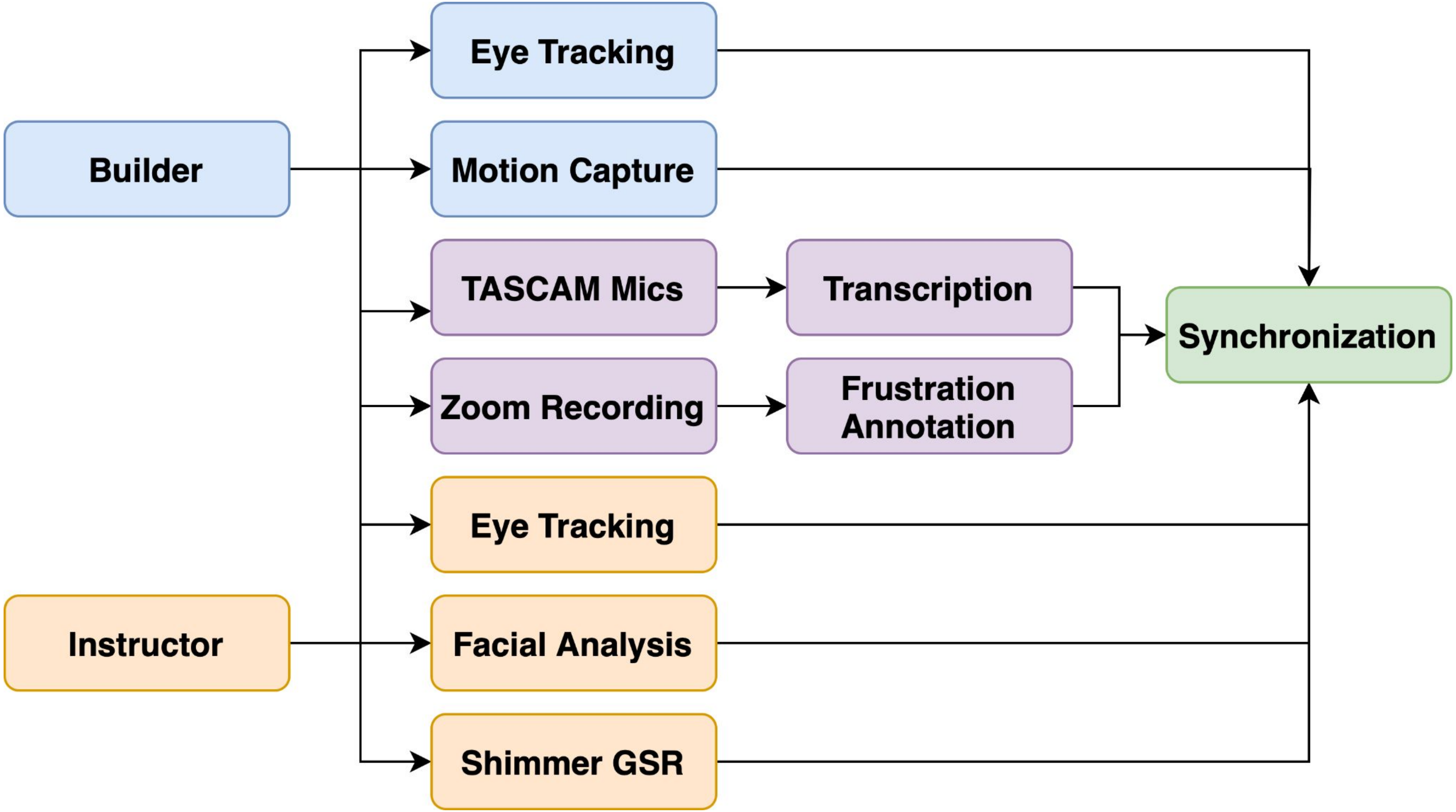


Common Modalities



Annotation

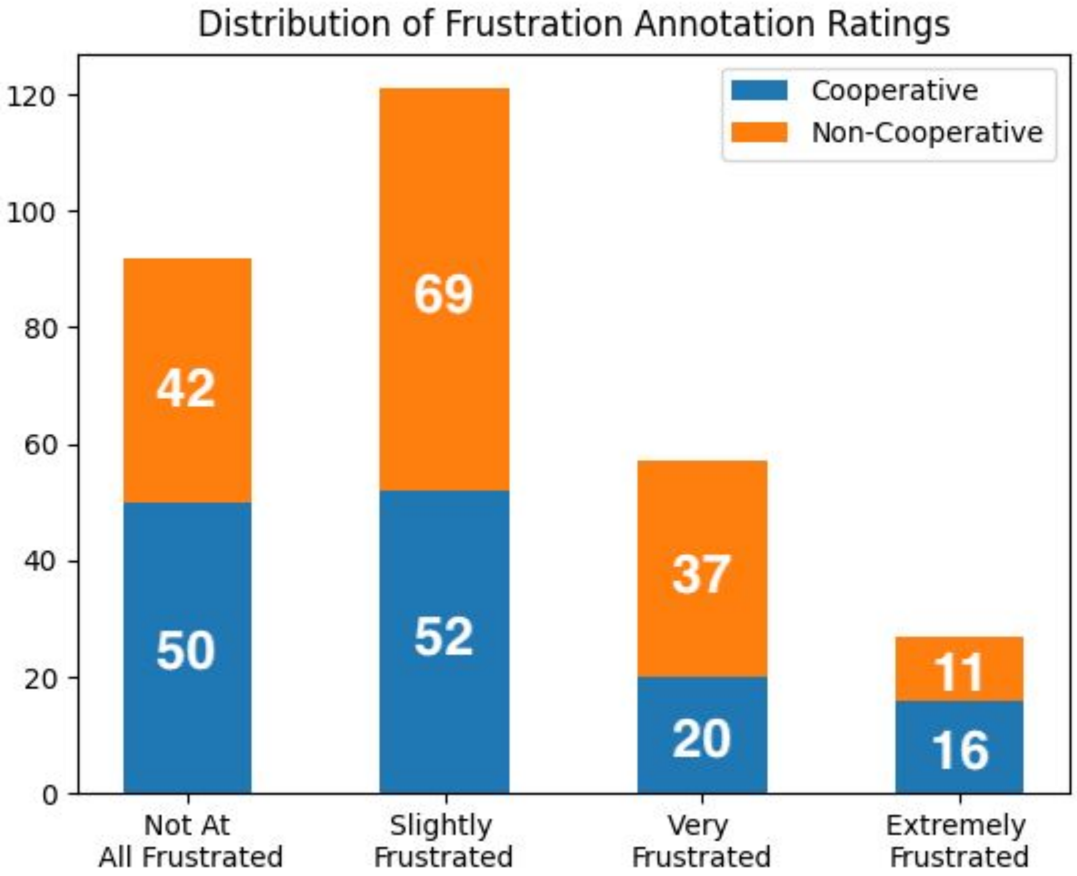

Michael Peechatt
Michael Peechatt (RIT Student)
Not At All Frustrated
Slightly Frustrated
Very Frustrated
Extremely Frustrated

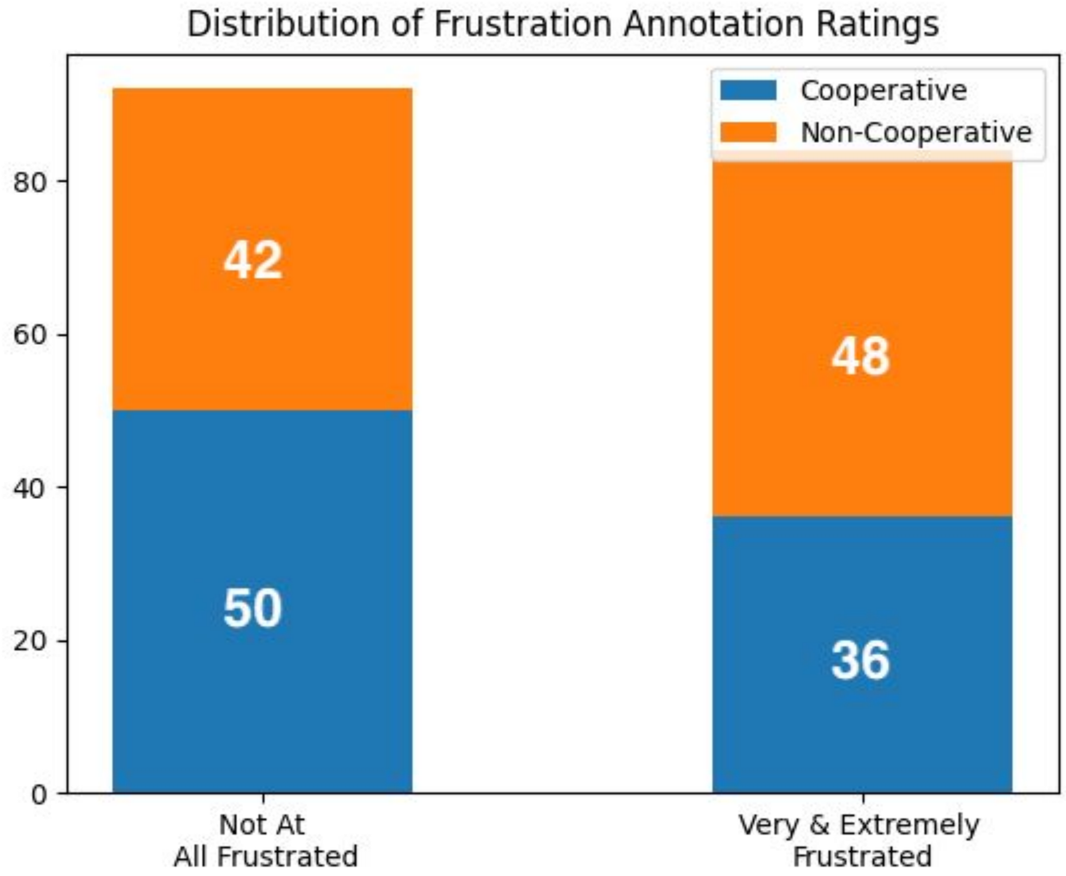

Builder
Instructor
Eye Tracking
Motion Capture
TASCAM Mics
Zoom Recording
Eye Tracking
Facial Analysis
Shimmer GSR
Transcription
Frustration Annotation
Synchronization

MULTICOLLAB Dataset

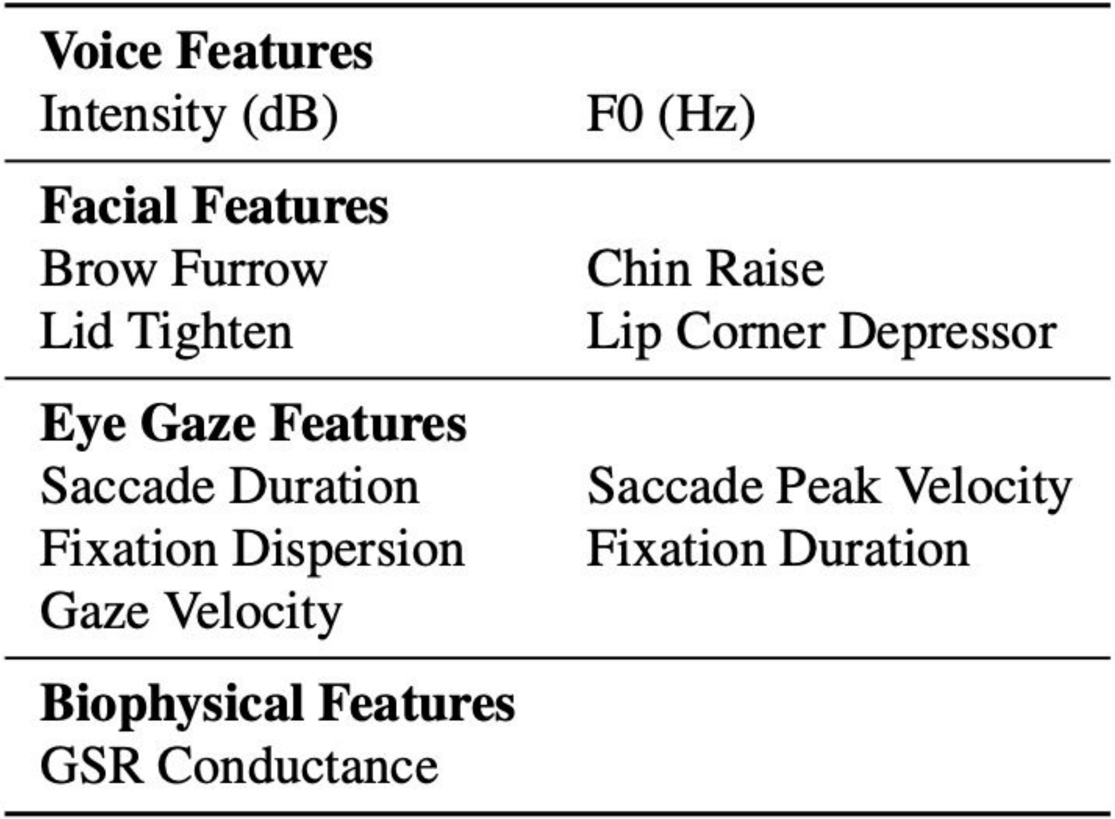
- ❑ **48 subjects (24 builder-instructor groups)**
  - ❑ 42% Female, 56% Male, 2% undisclosed
  - ❑ 8 groups had different gendered interactions
  - ❑ 16 groups had same gendered interactions

- ❑ **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

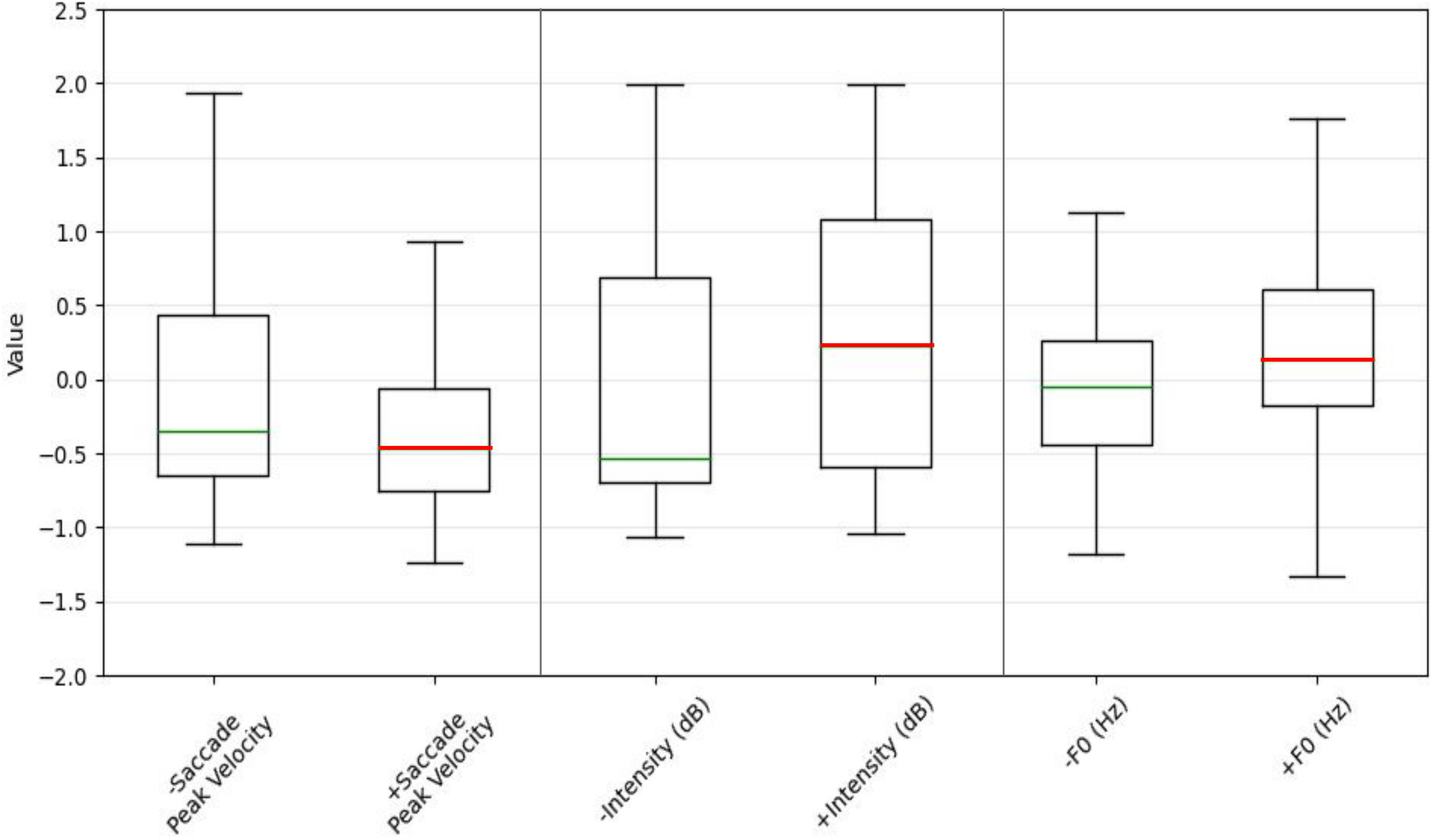

Distribution of Frustration Annotation Ratings
Cooperative
Non-Cooperative
0
20
40
60
80
100
120
42
50
69
52
37
20
11
16
Not At All Frustrated
Slightly Frustrated
Very Frustrated
Extremely Frustrated


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

Feature Extraction



Voice Features	
Intensity (dB)	F0 (Hz)
Facial Features	
Brow Furrow	Chin Raise
Lid Tighten	Lip Corner Depressor
Eye Gaze Features	
Saccade Duration	Saccade Peak Velocity
Fixation Dispersion	Fixation Duration
Gaze Velocity	
Biophysical Features	
GSR Conductance	


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

Dataset Shape

$$\begin{array}{cc} \begin{matrix} m_1 & m_2 & \cdots & m_n \end{matrix} & \\ \begin{bmatrix} & & t_1 & \\ & & t_2 & \\ & & \vdots & \\ & & t_n & \end{bmatrix} & \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{matrix} \end{array}$$

Dataset Shape

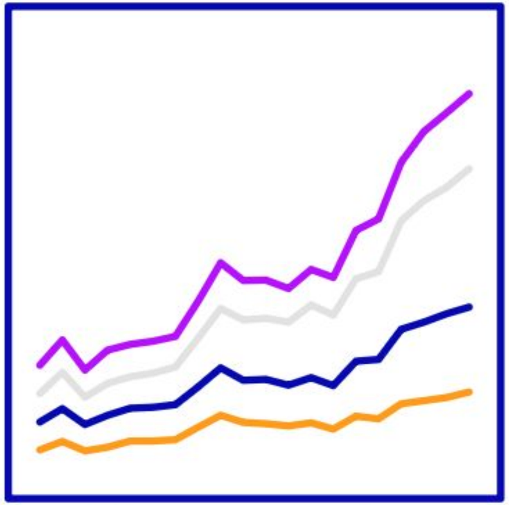
$$\begin{array}{cc}
\begin{array}{cccc} m_1 & m_2 & \cdots & m_n \end{array} & \\
\left[\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_n \end{array}\right] & \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_n \end{array}
\end{array}$$

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

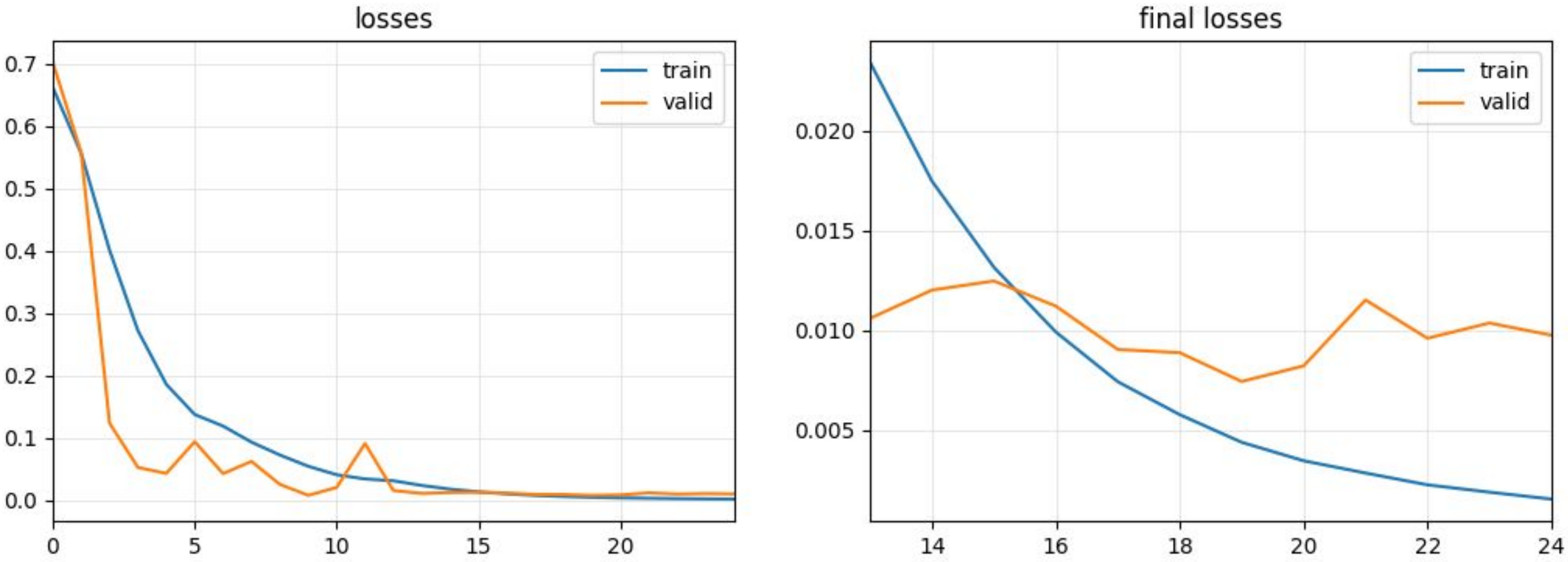
$$||\{m_1, m_2, \ldots, m_n\}|| \cdot t_n$$

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

tsai



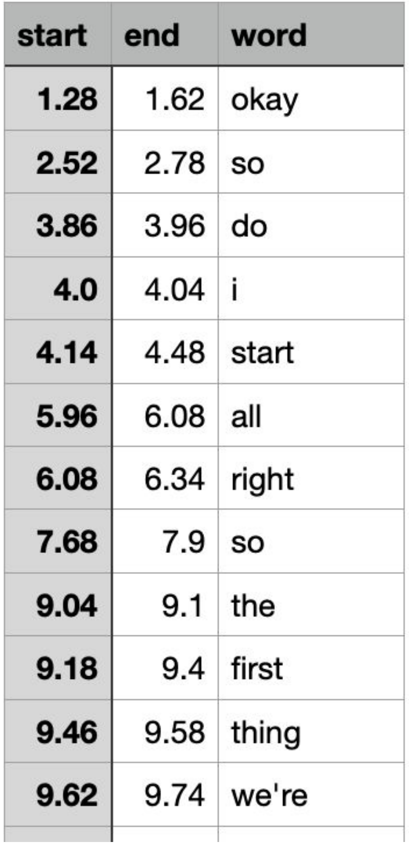
state-of-the-art deep learning for time series and sequences



t_win = 10000 milliseconds, t_n = 20

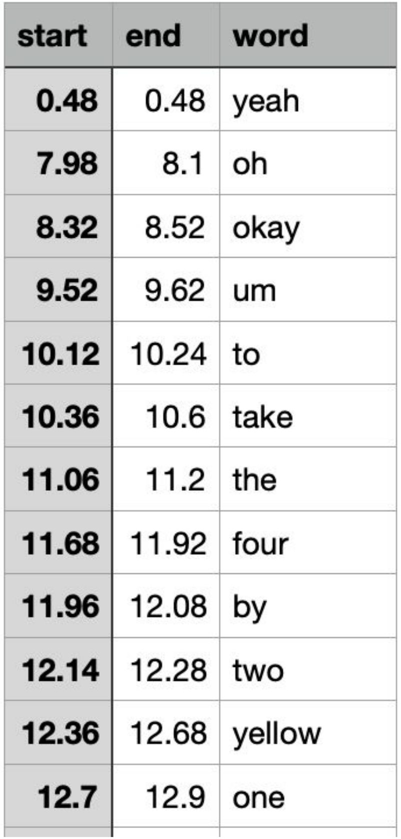
avg. accuracy = 0.523

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

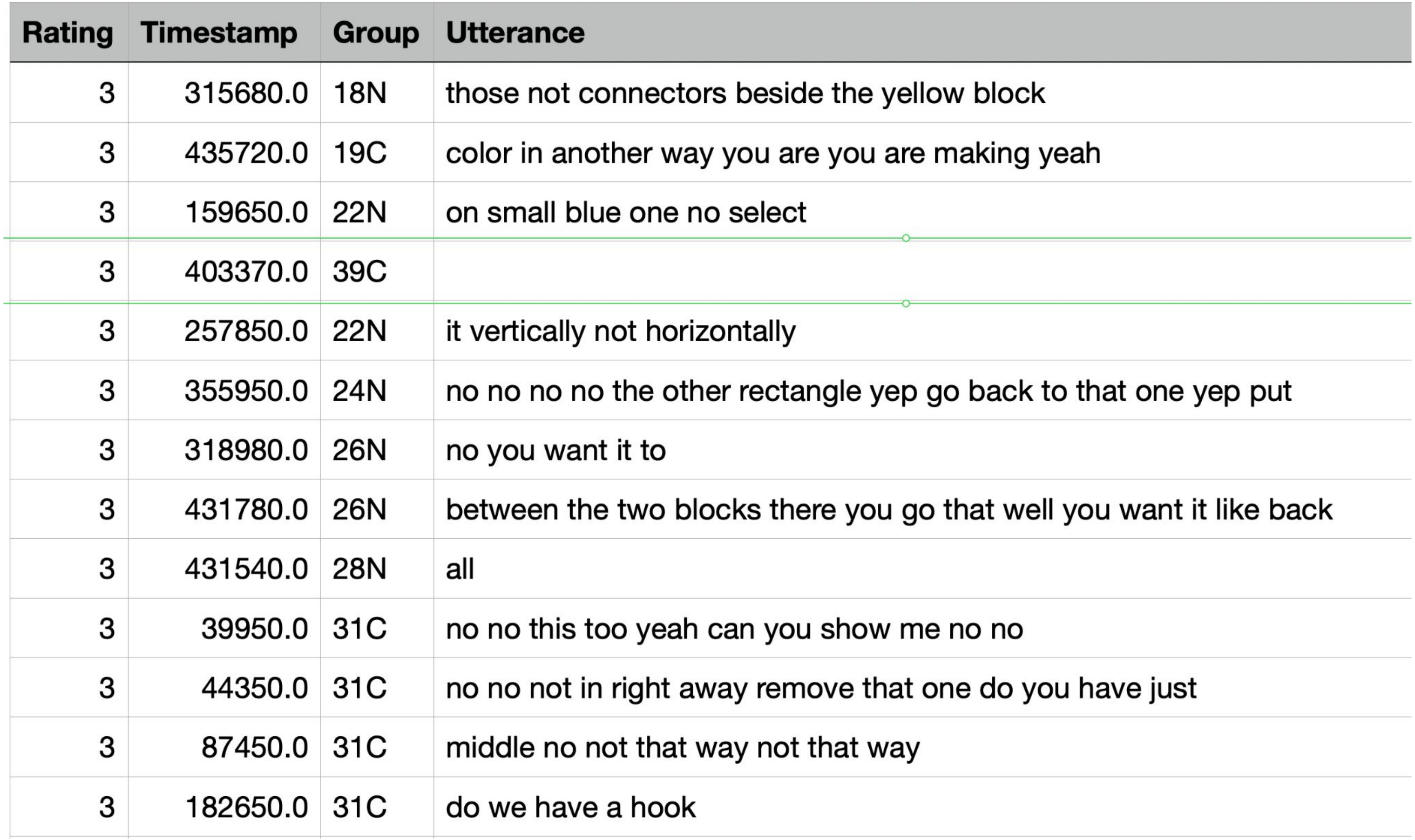
27CI_transcript



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

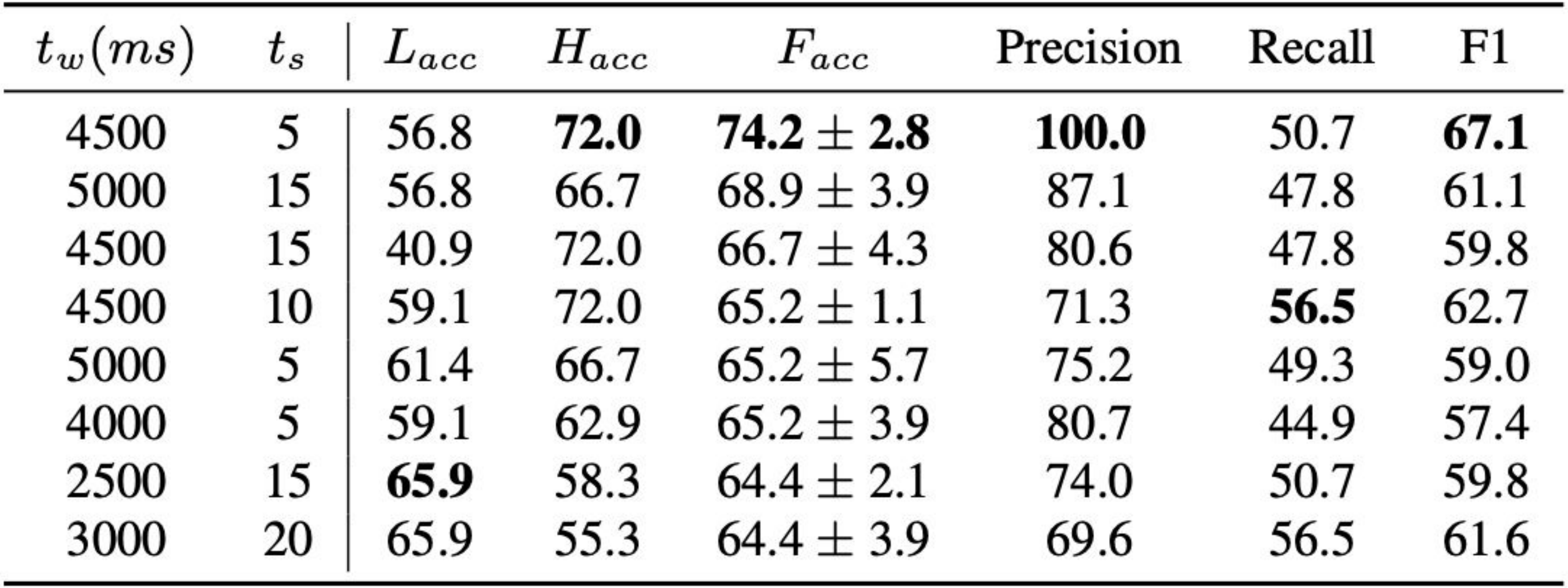
fastText

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



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



$t_w(ms)$	t_s	L_{acc}	H_{acc}	F_{acc}	Precision	Recall	F1
4500	5	56.8	**72.0**	**74.2 ± 2.8**	**100.0**	50.7	**67.1**
5000	15	56.8	66.7	68.9 ± 3.9	87.1	47.8	61.1
4500	15	40.9	72.0	66.7 ± 4.3	80.6	47.8	59.8
4500	10	59.1	72.0	65.2 ± 1.1	71.3	**56.5**	62.7
5000	5	61.4	66.7	65.2 ± 5.7	75.2	49.3	59.0
4000	5	59.1	62.9	65.2 ± 3.9	80.7	44.9	57.4
2500	15	**65.9**	58.3	64.4 ± 2.1	74.0	50.7	59.8
3000	20	65.9	55.3	64.4 ± 3.9	69.6	56.5	61.6

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?