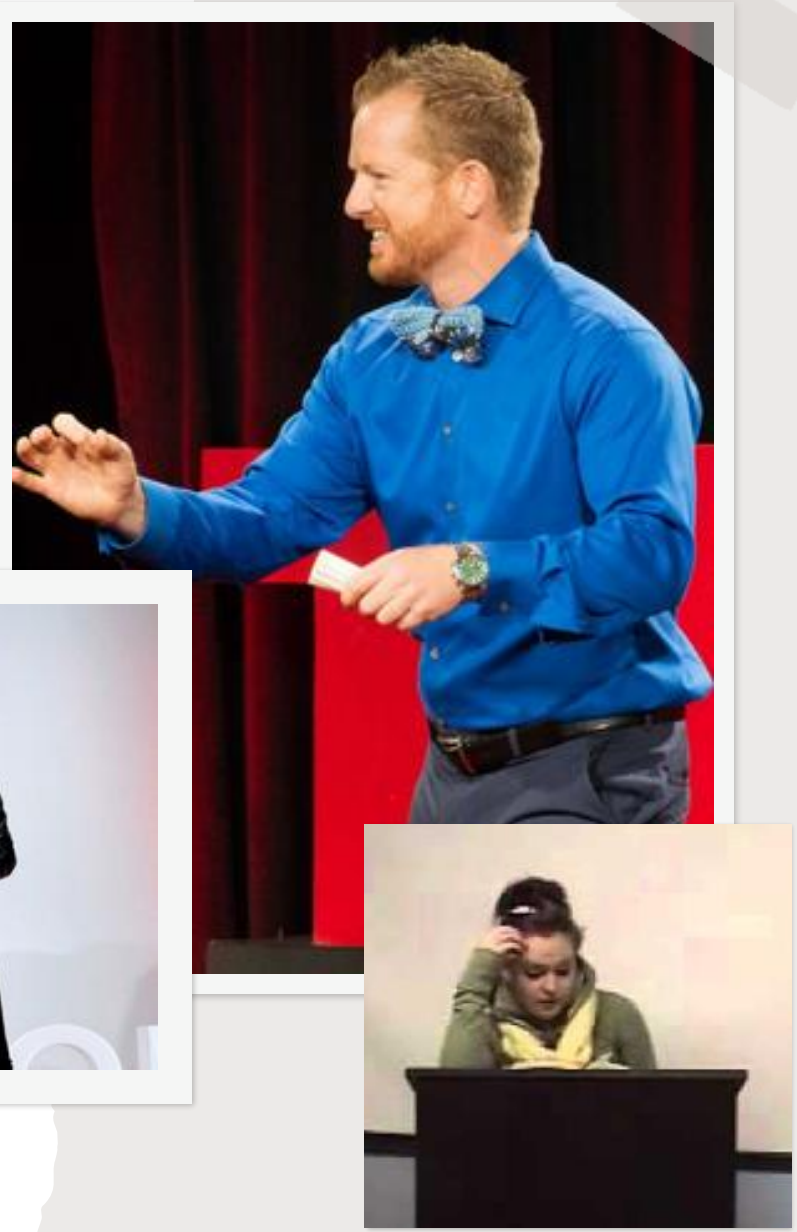
CAPSTONE
PUBLIC SPEAKER ANALYZER

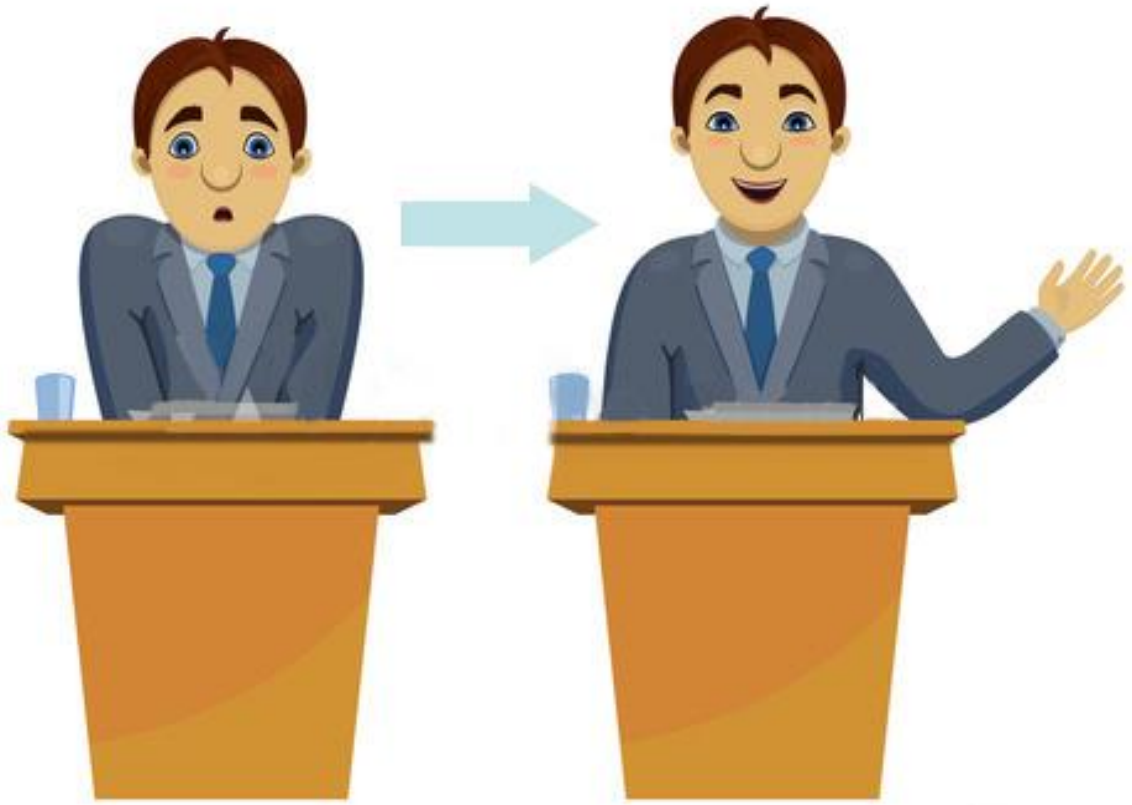
GUIDED BY: SUDARSHAN IYENGAR

SUBMITTED BY:
KUSHAL AGRAWAL - 2020CSB1096
TUMMALA SANJANA REDDY – 2020CSB1137



PROBLEM STATEMENT

Identify and classify speakers as "**good**" or "**bad**"

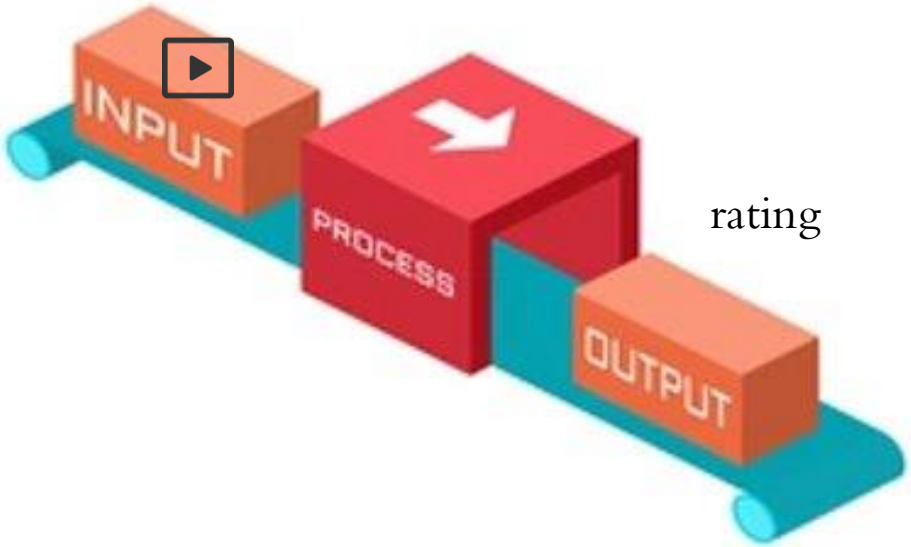


ATTRIBUTES CONSIDERED

- Emotions
- Pose Landmarks
- Hand gestures
- Head Movement
- Words selection (content)

INPUT AND OUTPUT

- **Input:** A video featuring a speaker delivering a speech.
- **Output**: Rating the speaker based on attributes.



OBJECTIVES

- Finding datasets
- Curating Data Sets
- Feature Extraction

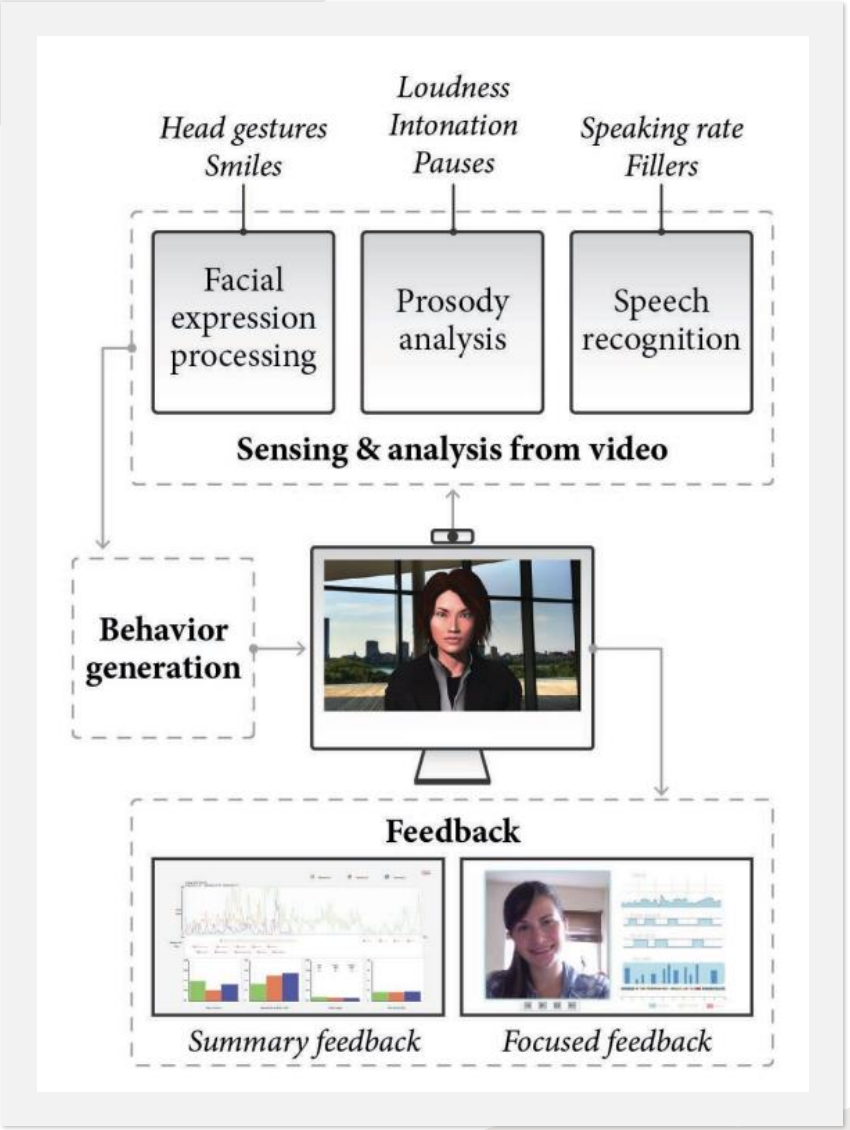
OBJECTIVES

- Neural Network Implementation
- Training
- Testing Accuracy

RESEARCH WORK

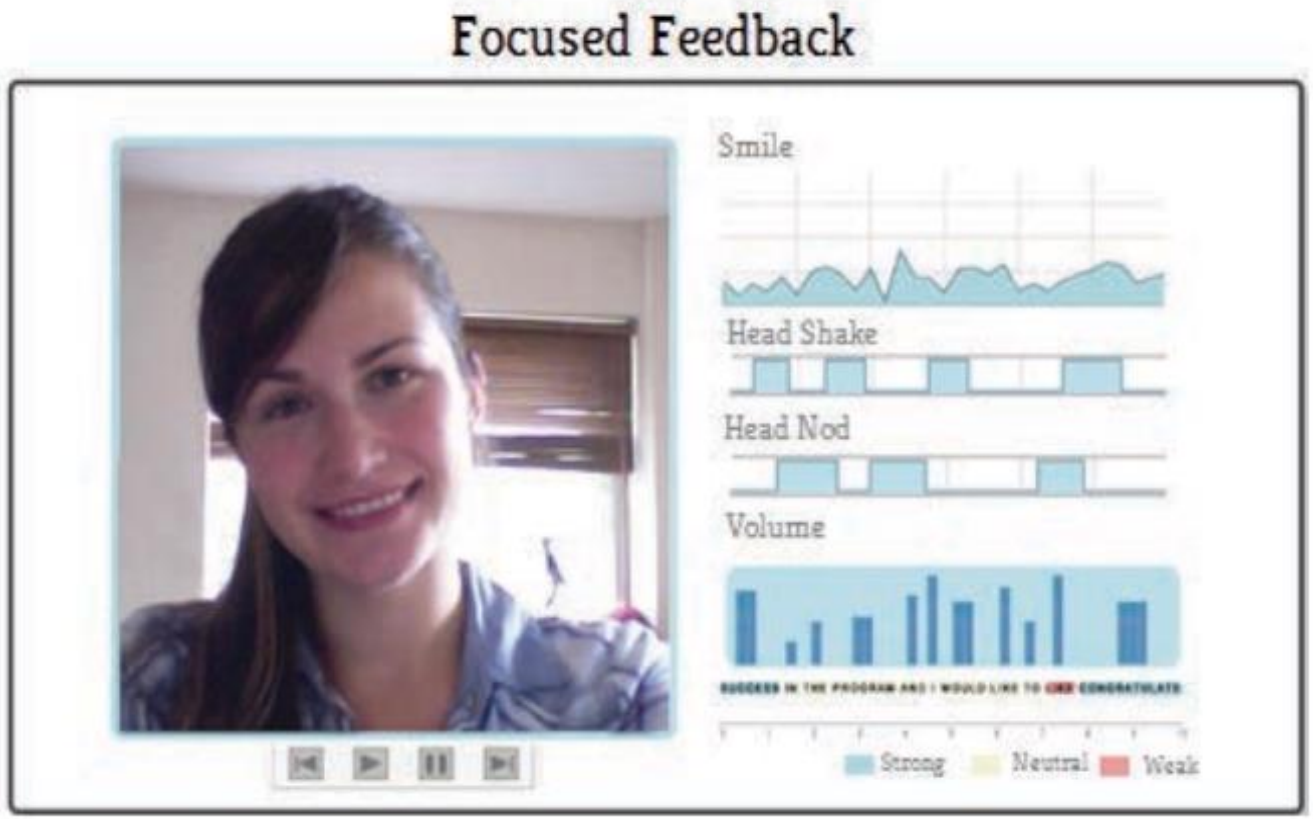
MACH: MY AUTOMATED CONVERSATION COACH

- MACH poses interview questions and observes user **behavior**.
- Used for job **interview** training.



MACH: MY AUTOMATED CONVERSATION COACH

- Analyzes **facial expressions** and speech, generating behaviors.
- Offers visual **feedback** on user performance after each interaction.



BODILY BEHAVIORS IN SOCIAL INTERACTION: NOVEL ANNOTATIONS AND STATE-OF-THE-ART EVALUATION

- Introducing **Bodily Behaviors in Social Interaction** (BBSI) annotations.

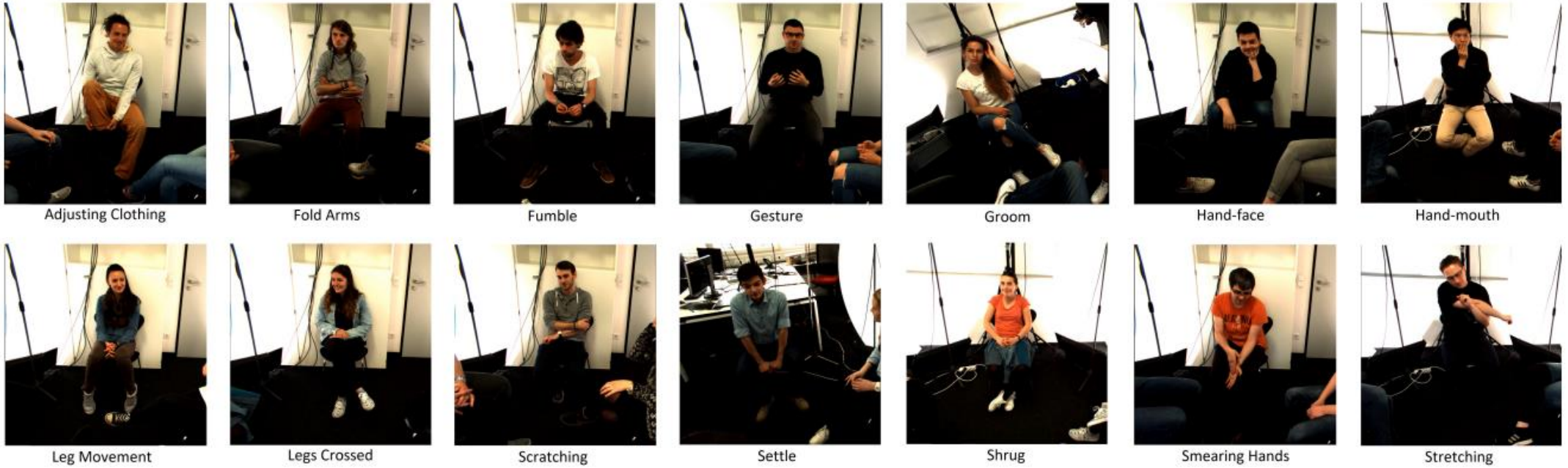
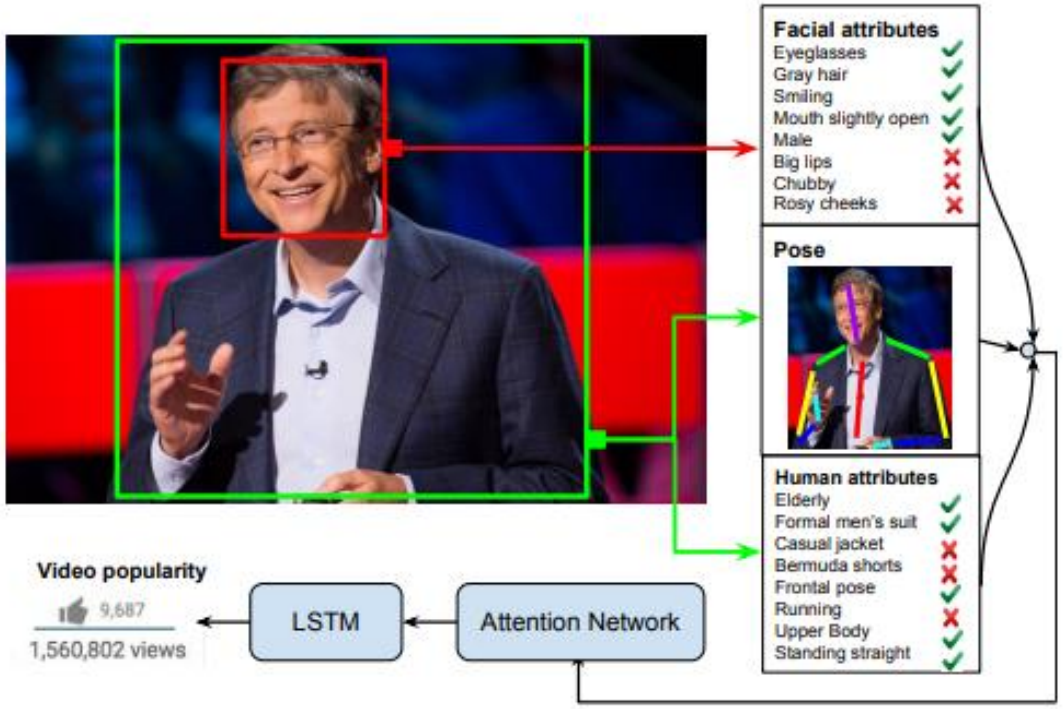


Figure 1: Examples of annotated bodily behaviors.

MULTICHANNEL ATTENTION NETWORK FOR ANALYZING VISUAL BEHAVIOR IN PUBLIC SPEAKING

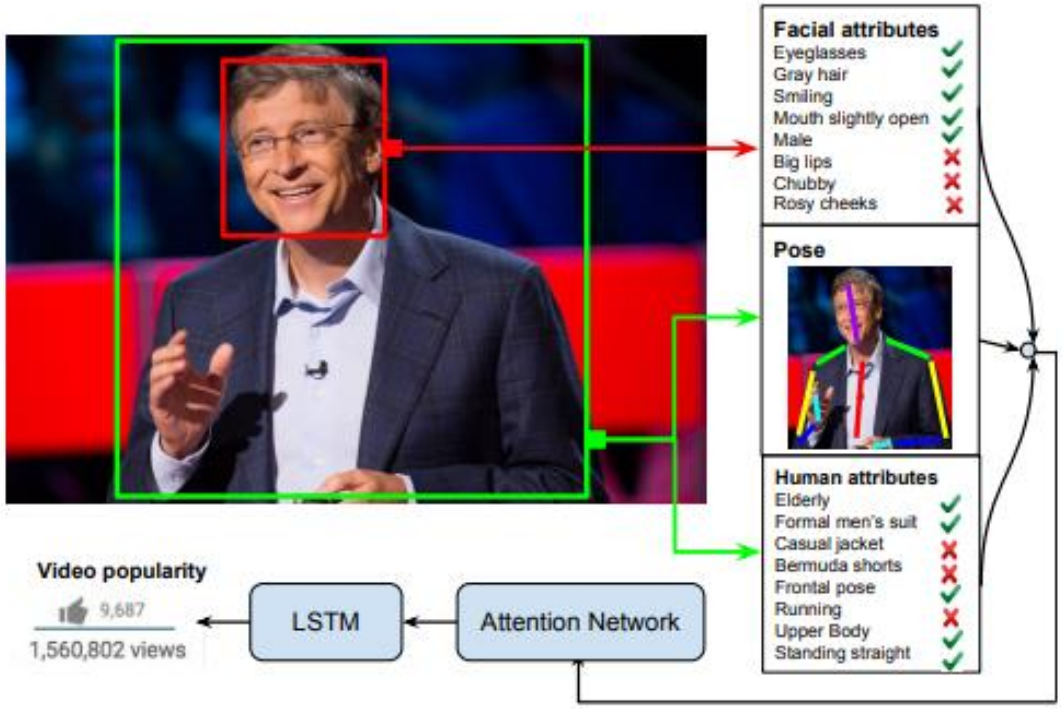
Visual cues related to facial and physical appearance, facial expressions, and pose variations are learned using convolutional neural networks (CNN) connected to an attention-based long short-term memory (LSTM) network to predict the video popularity



MULTICHANNEL ATTENTION NETWORK FOR ANALYZING VISUAL BEHAVIOR IN PUBLIC SPEAKING

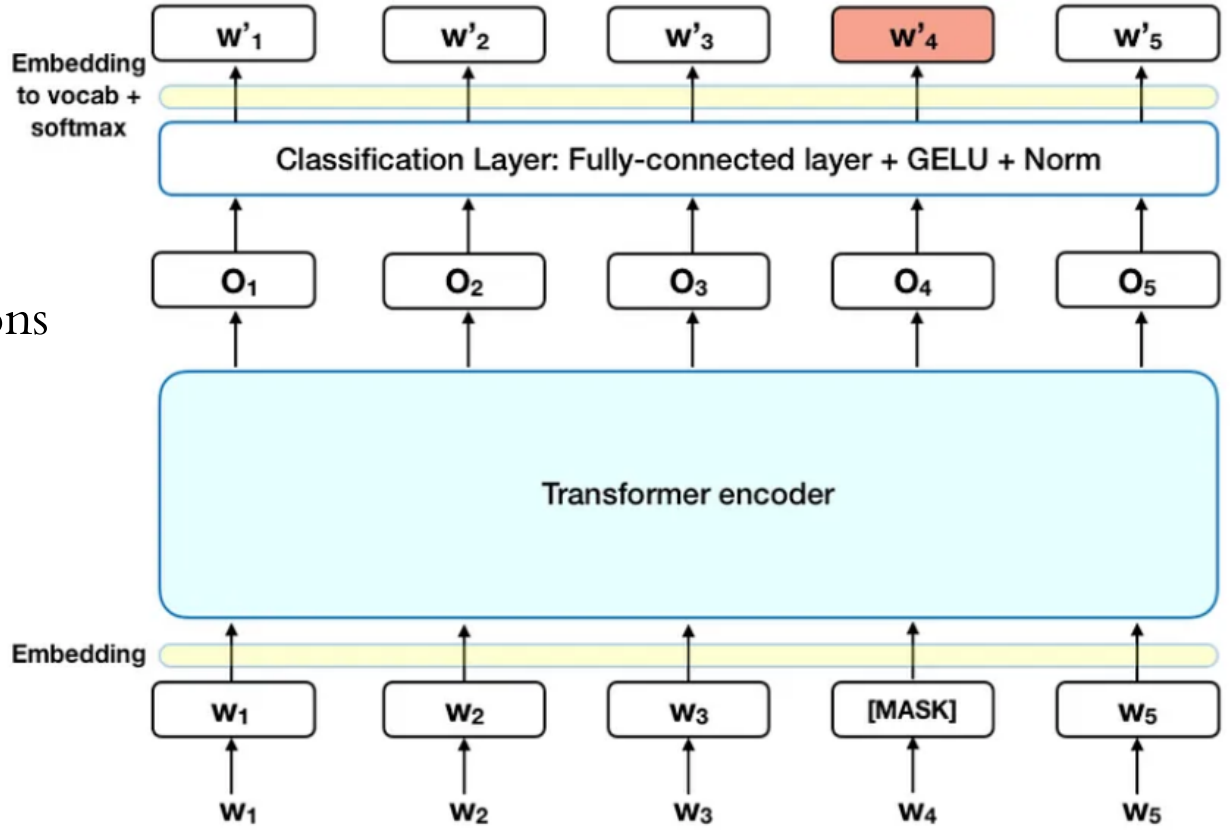
Factors not considered-

- Gestures
- Emotions expressed
- Content of speech
- Voice modulation



BERT MODEL

- Bidirectional Encoder Representations from Transformers



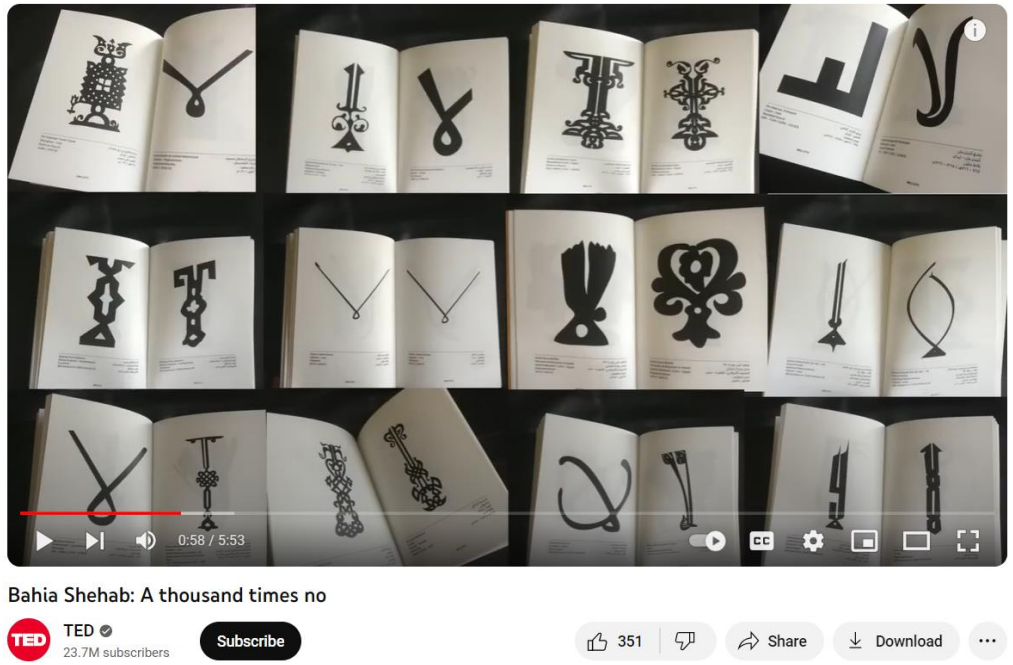
WORK DONE

FINDING DATA SETS

https://www.kaggle.com/datasets/jeniagerasimov/ted-talks-info-dataset/

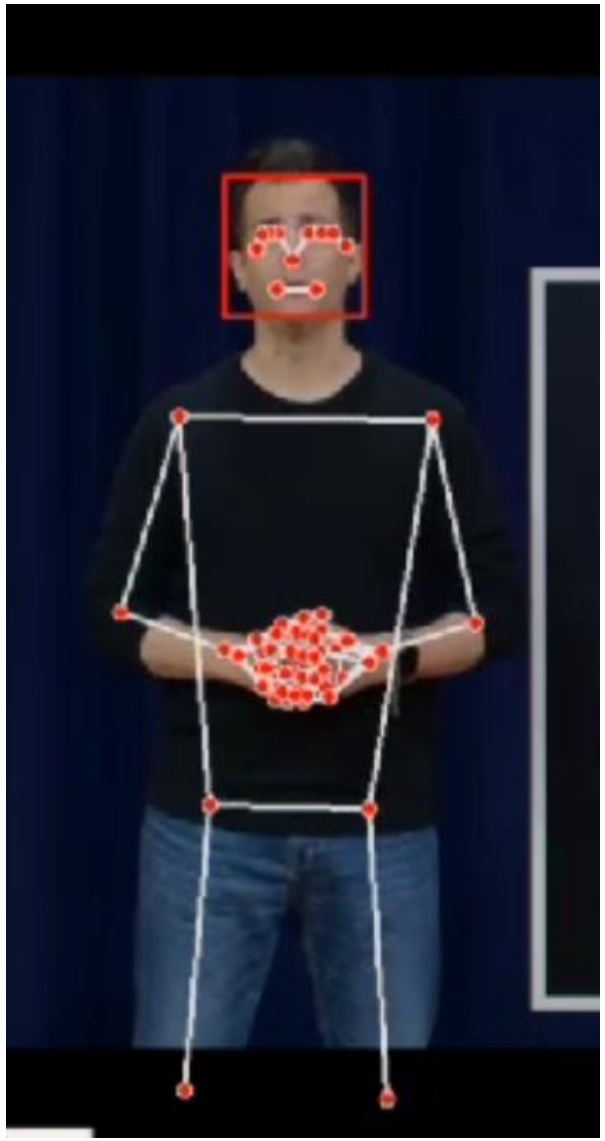
CURATING DATA SETS

- Removing educational videos.
- Videos where speaker is not visible at all.



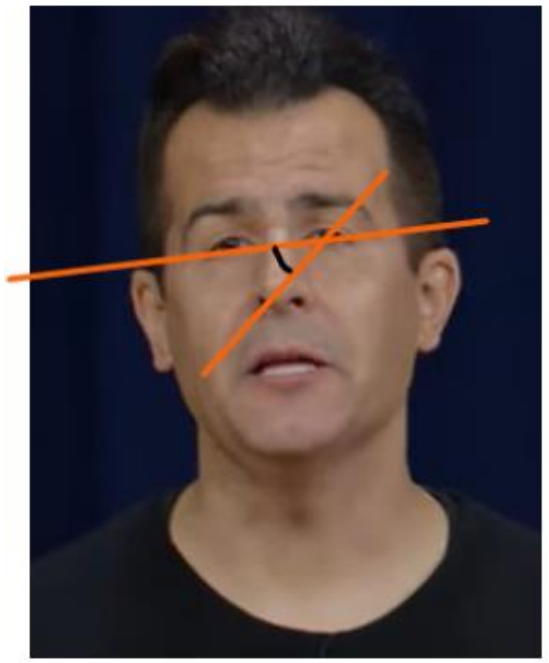
FEATURE EXTRACTION

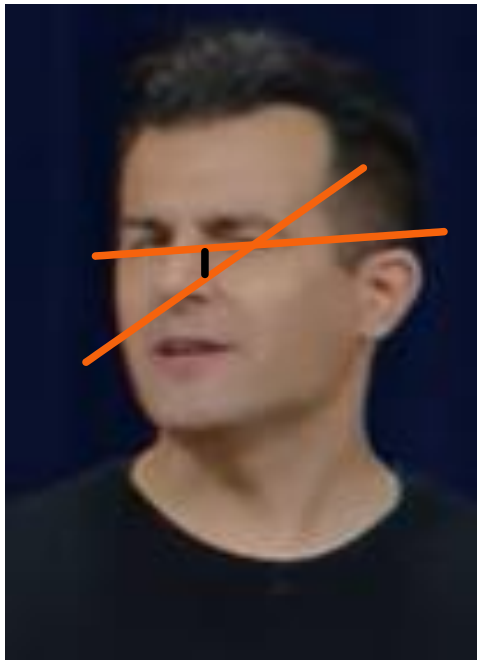
- Hand gestures
- Pose features



FEATURE EXTRACTION

- Emotions
- Head turn angles (denoting head movements)





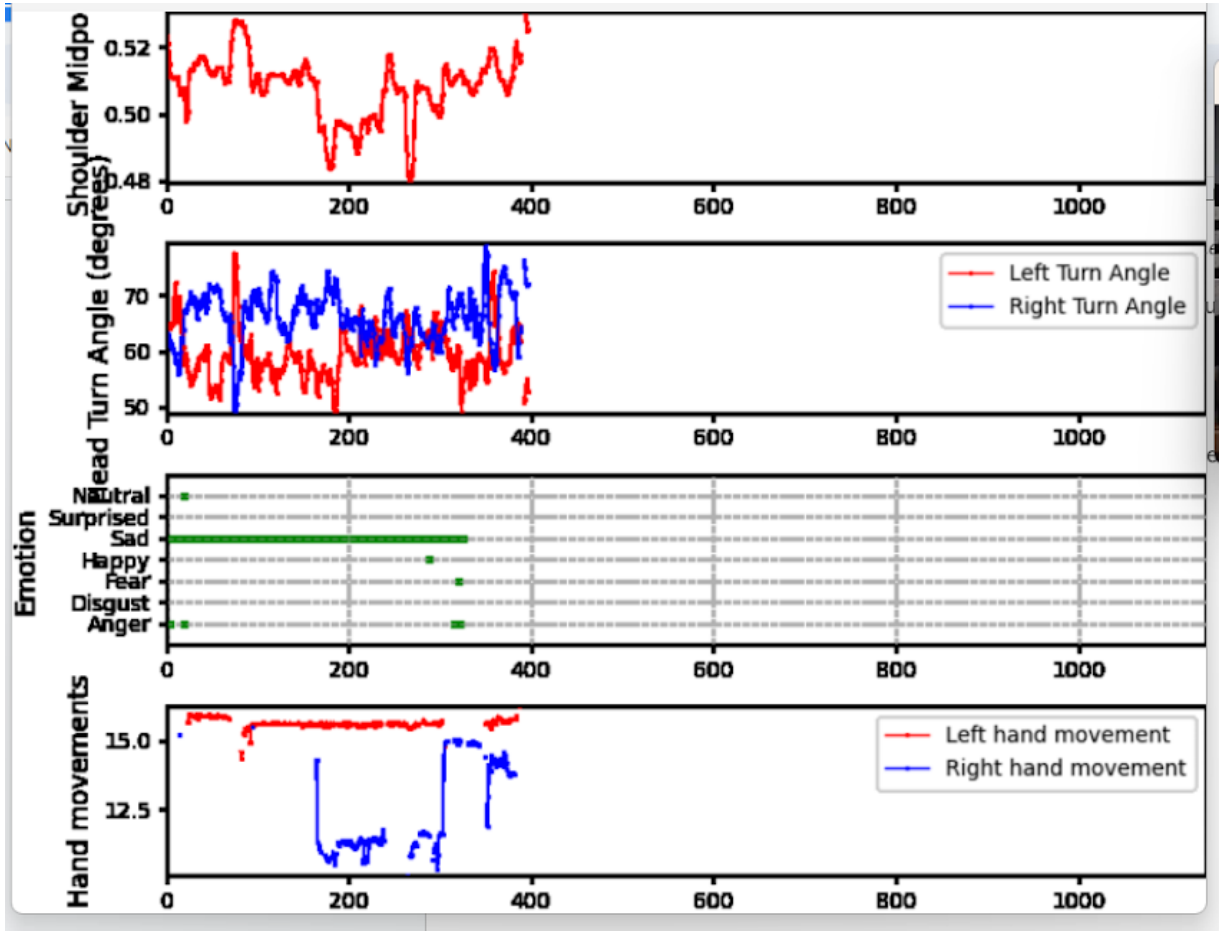
FEATURE EXTRACTION

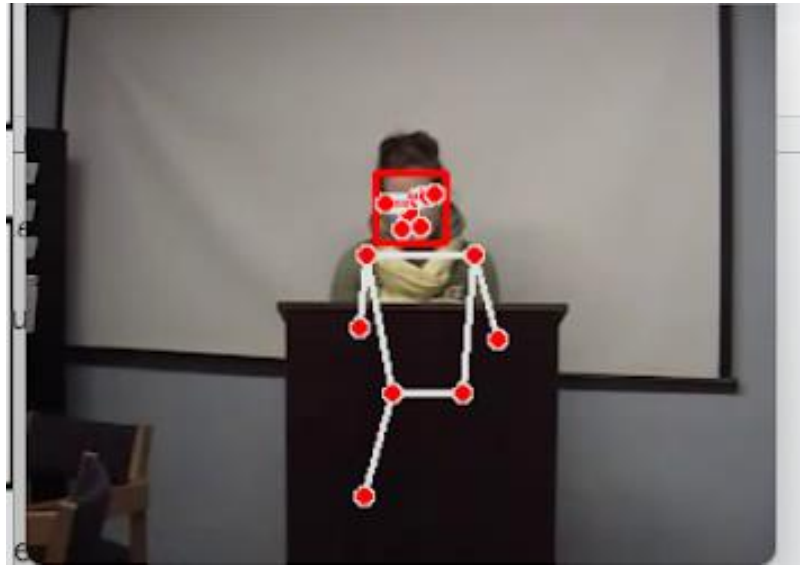
- Words chosen : Transcripts

avgWordsSpoken_t1_t2 =ceil[(total words spoken at timestamp t1)/ (t2 -t1)]

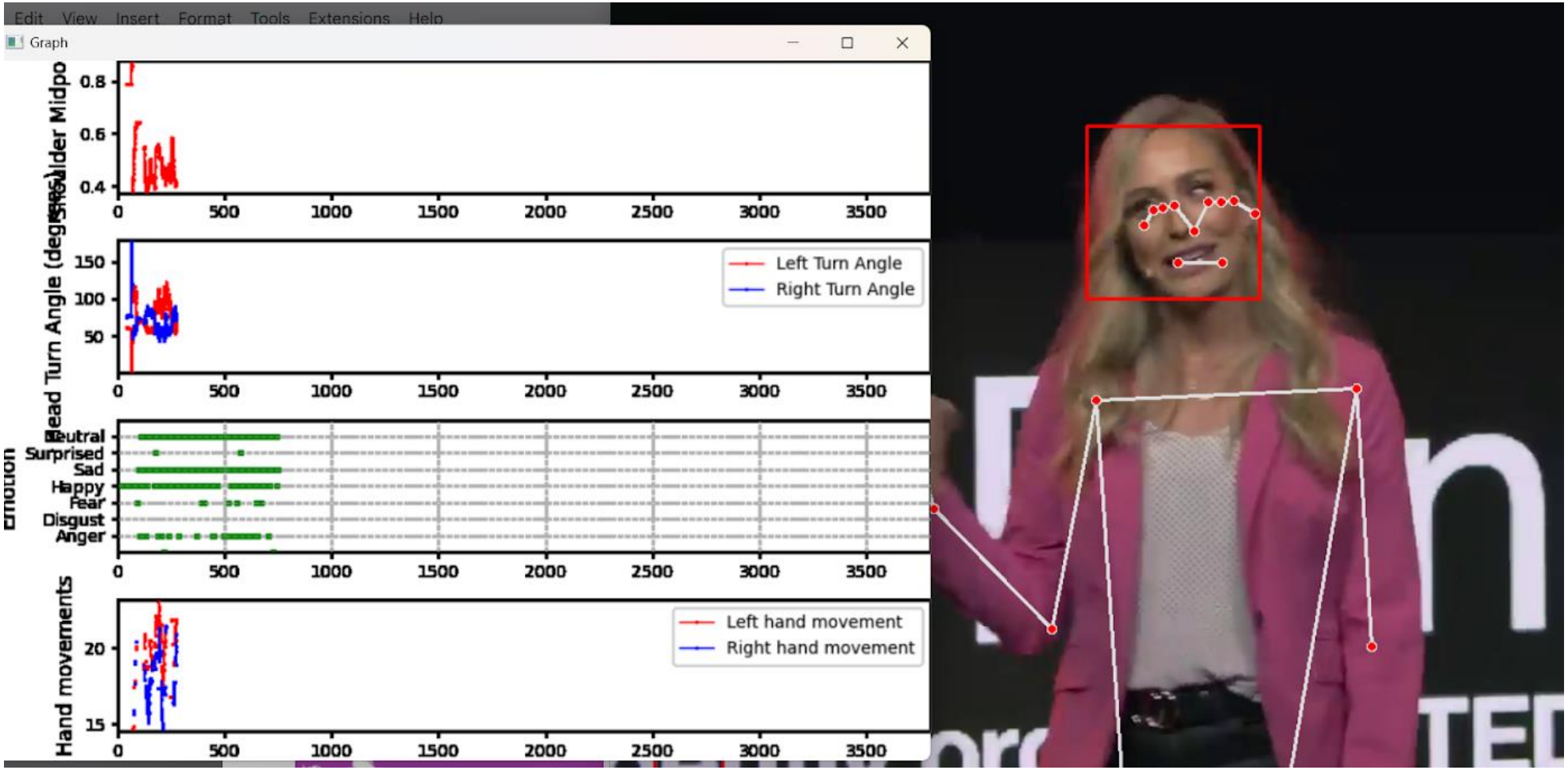
```
[16.26] First of all, I'm a geek.
[19.26] I'm an organic food-eating,
[21.26] carbon footprint-minimizing, robotic surgery geek.
[24.26] And I really want to build green,
[27.26] but I'm very suspicious
[29.26] of all of these well-meaning articles,
[31.26] people long on moral authority
[33.26] and short on data,
[35.26] telling me how to do these kinds of things.
[37.26] And so I have to figure this out for myself.
[39.26] For example: Is this evil?
[42.26] I have dropped a blob of organic yogurt
[45.26] from happy self-actualized local cows
```

DASHBOARD : OBSERVING CORRELATIONS

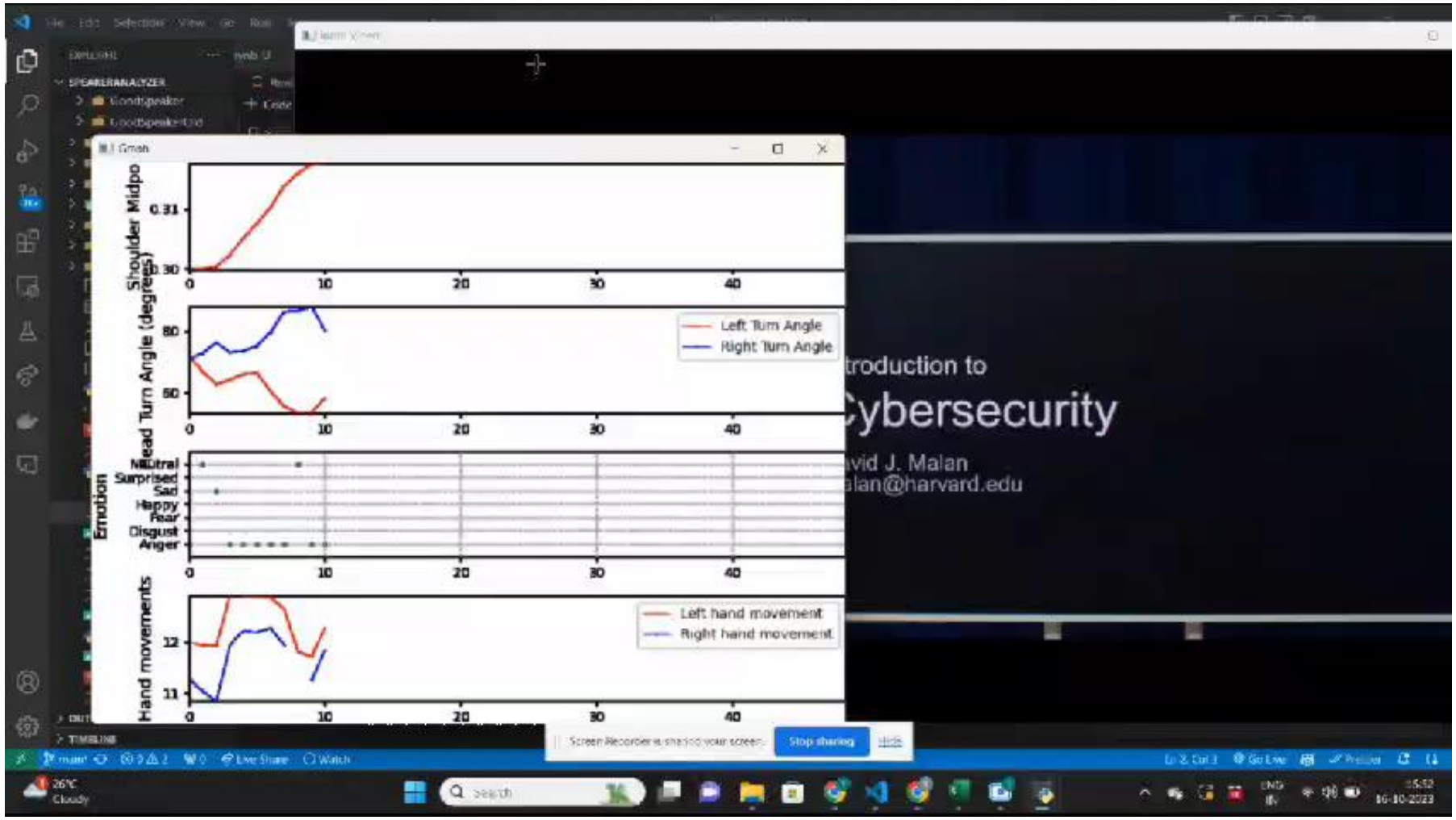




DASHBOARD : OBSERVING CORRELATIONS



DASHBOARD : OBSERVING CORRELATIONS



EXTRACTING FEATURES PER SECOND

The features matrix for each second is represented as follows -

- Emotion:
  - One-hot-encoding representing one of 7 emotions
  - Neutral, Surprised, Sad, Happy, Fear, Disgust, Anger
  - Size = 7

EXTRACTING FEATURES PER SECOND

- Body center:
  - x coordinates of shoulder midpoint
  - Size = 1

EXTRACTING FEATURES PER SECOND

- Head turn angle:
  - left eye angle and right eye angle
  - Size = 2

EXTRACTING FEATURES PER SECOND

- Pose landmarks:
  - Coordinates of 33 pose landmarks
  - Size = 66 (both x, y coordinates)

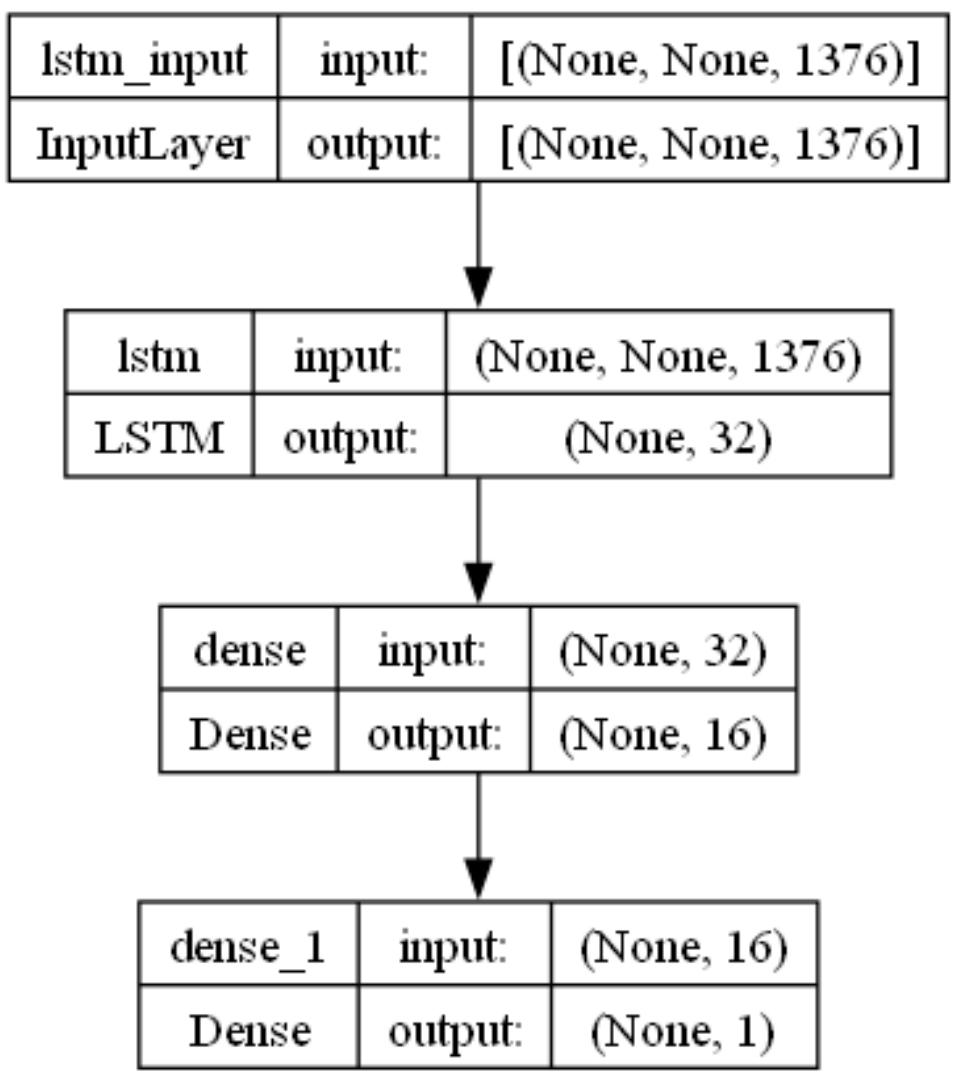
All extracted for 8 frames in a second

EXTRACTING FEATURES PER SECOND

- Text embedding
  - Embedding of content spoken in each second
  - Size = 768 (obtained using BERT mode)

Total size of matrix = (7+1+2+66)*8+768 = 1376

NEURAL NETWORKS



TRAINING

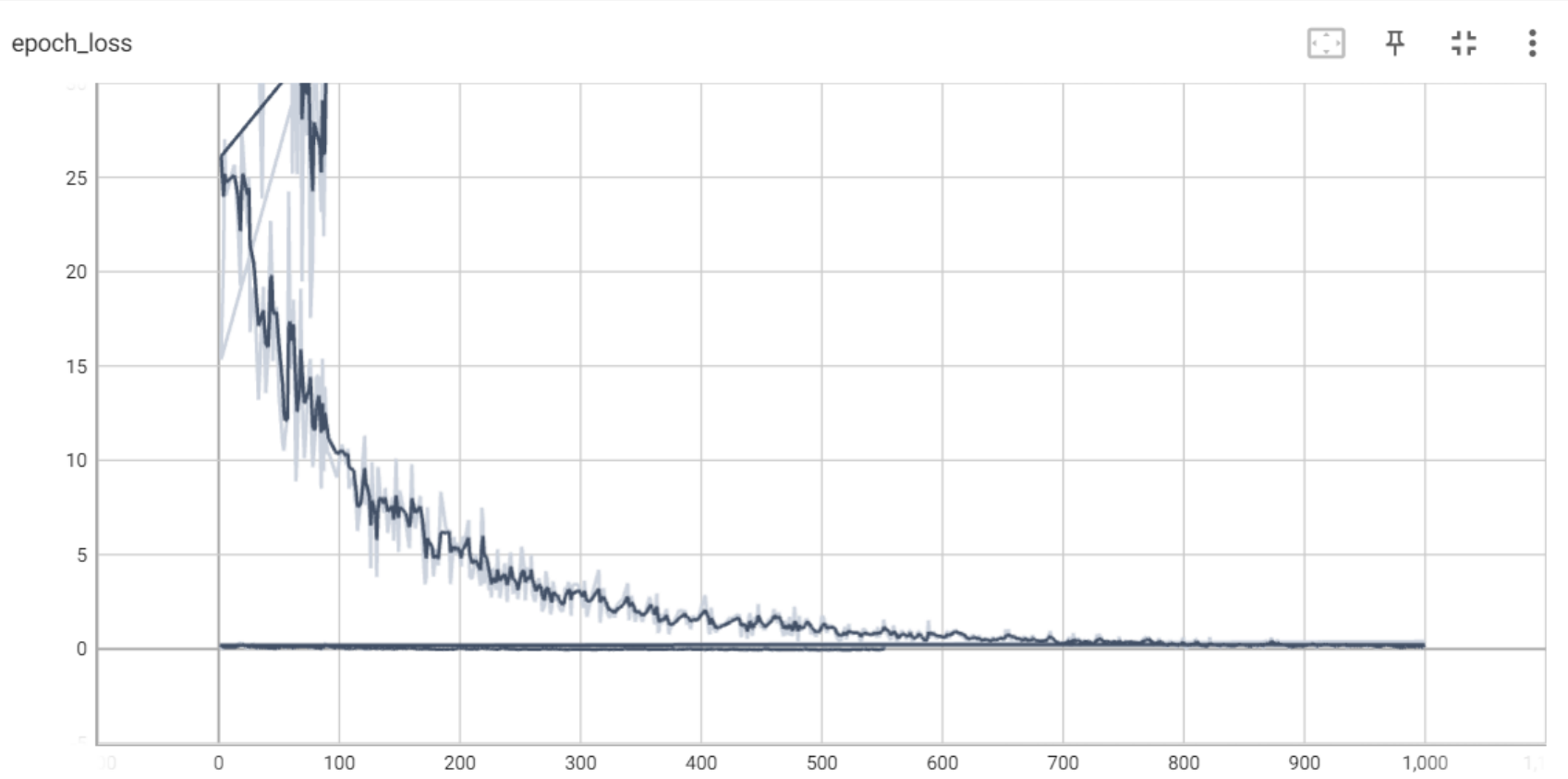


Fig: Learning curve (plot of loss values with epoch)

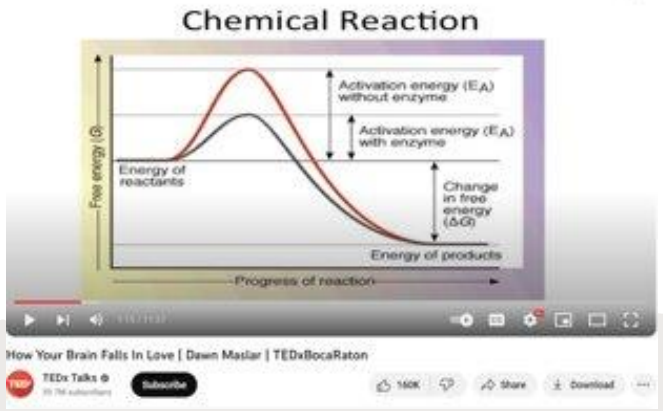
RESULTS ON TESTING DATA

```
1/1 [==============================] - 0s 280ms/step - loss: 0.0058
Test Loss: 0.005760721862316132
1/1 [==============================] - 0s 358ms/step
Test Mean Absolute Error: 0.02238837075417429
Video ID:  1232
Window ID: 1
Prediction: 0.029358020052313805, Original Label: [0.02942262]


Window ID: 2
Prediction: 0.029358020052313805, Original Label: [0.02942262]


Window ID: 3
Prediction: -0.269319623708725, Original Label: [0.02942262]


Window ID: 4
Prediction: -0.024634459987282753, Original Label: [0.02942262]
```

```
Video ID:  1537
Window ID: 1
Prediction: 0.029358020052313805, Original Label: [0.02966094]


Window ID: 2
Prediction: 0.029358020052313805, Original Label: [0.02966094]


Window ID: 3
Prediction: 0.029358020052313805, Original Label: [0.02966094]


Window ID: 4
Prediction: 0.029358020052313805, Original Label: [0.02966094]
```

CHALLENGES

- Curating Data Set
- Speaker Invisible in some parts
- Multiple speakers



FUTURE WORK & IMPROVEMENTS

- Audio features adds more meaning to transcripts.
- Words based on lip movements rather than average.

FUTURE WORK & IMPROVEMENTS

- Work on the existing neural network to increase accuracy.
- Experimenting on other transformers and choosing the best one.

THANK YOU!