Multimodal Emotion Recognition

Final Project Discussion

19D180005 Akshat Vira
200010006 Ajinkya Patil
22M1072 Prashant Khatri
22M1079 Mayank Pershad

Problem Statement

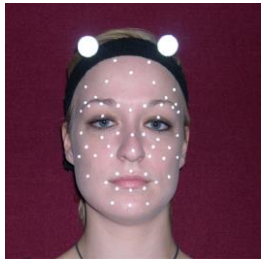
Our objective is to develop a **robust multimodal emotion recognition model** that leverages **acoustic and lexical features**. We aim to **investigate attention mechanisms** for extracting **emotionally relevant temporal aggregates** from the sequential model blocks and explore a modality-level attention model for **fusing lexical and acoustic** information effectively. This work will contribute to improved performance in emotion recognition and enhance human-computer interaction systems.

Related Work

- Björn W. Schuller. 2018. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61, 5 (May 2018), 90–99. https://doi.org/10.1145/3129340
- S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952552. LINK
- Kun Han & Yu, Dong & Tashev, Ivan. (2014). Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. LINK
- Gustavo & Rozgic, Viktor & Wang, Weiran & Wang, Chao. (2019). Multimodal and Multi-view Models for Emotion Recognition. LINK
- Arevalo, John & Solorio, Thamar & Montes, Manuel & González, Fabio. (2017). Gated Multimodal Units for Information Fusion. LINK
- Neural Machine Translation by Jointly Learning to Align and Translate Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio LINK

Dataset: IEMOCAP

1. **12 hours** of **scripted or improvised dyadic** conversation with 10 actors (**5 female, 5 male**). (English)
2. **Available Modalities: Speech**, video, motion capture data, **dialog Transcriptions, Word level, Syllable level, and Phoneme level alignment.**
3. **Categorical and dimensional labels**
4. Manual segmentation and annotation by multiple **human annotators (>3)**



Workflow and Technique

Overview of our approach:

1. Data Extraction
2. Data Pre-Processing
3. Feature Extraction
4. Splitting data into test and Train(20-80 split, Speaker Dependent split)
5. Model Architecture
6. Model training and testing
7. Result Analysis
8. Compile best performing models and create a demo

Workflow and Technique: Data Preprocessing & Feature Extraction

Processing

1. Removed 2 samples that have the audio file but not the transcribed file
2. Removed the samples with rater agreement = 0 (i.e, no rating overlap)
3. unbalanced data issue: Removed class categories with insufficient samples (in the end, we have 4 categories: anger, happiness, neutral, and sadness with a total of ~5530 samples)
4. Merged the sample from the happiness and excitement emotion category

Emotion	Sample Count	Emotion	Sample Count
xxx	2506	fru	1849
neu	1708	ang	1103
sad	1084	exc	1040
hap	595	sur	107
fea	40	oth	3
dis	2		

Emotion	Sample Count
neu	1708
hap/excit	1635
ang	1103
sad	1084

Speaker	# Samples (M)	Speaker	# Samples (F)
M1	578	F1	507
M2	518	F2	505
M3	614	F3	536
M4	478	F4	553
M5	597	F5	644

Workflow and Technique: Data Preprocessing & Feature Extraction

Preprocessing cont..

5. Removed the samples that have only “++BREATHING++” or “++GARBAGE++” in the transcription. Samples with only “++LAUGHTER++” were not removed.
6. Removed “++GARBAGE++” from the transcriptions.
7. Used Contractions on the transcribed to change strings like That’s to That is, I’ve to I have, etc
8. used additional token for <sil> (silence), ++BREATHING++ and ++LAUGHTER++

The data has been processed and stored in the `‘data_processed.pickle’` in the shared drive link

Features:

- Acoustic: **`OpenSmile ComPare_2016` LLD** features (65 dimensions) and then **z-standardization** was performed by using mean and standard deviation of the test set features
- Lexical: **`BERT` word-level embedding** (768 dimensions)

Implemented the acoustic and the lexical-only models with average pooling and attention and compared the result

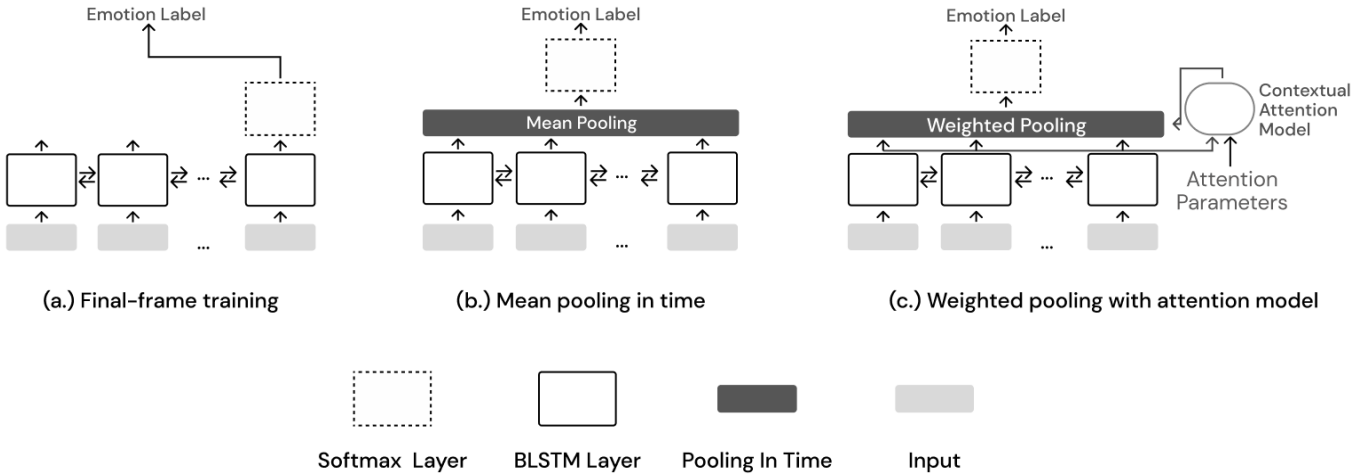
Architecture: Unimodal

Context-based attention Bahdanau et al. (2014).

$$e_i = v^\intercal \tanh(\mathrm{W}_h h_i + b_h)$$

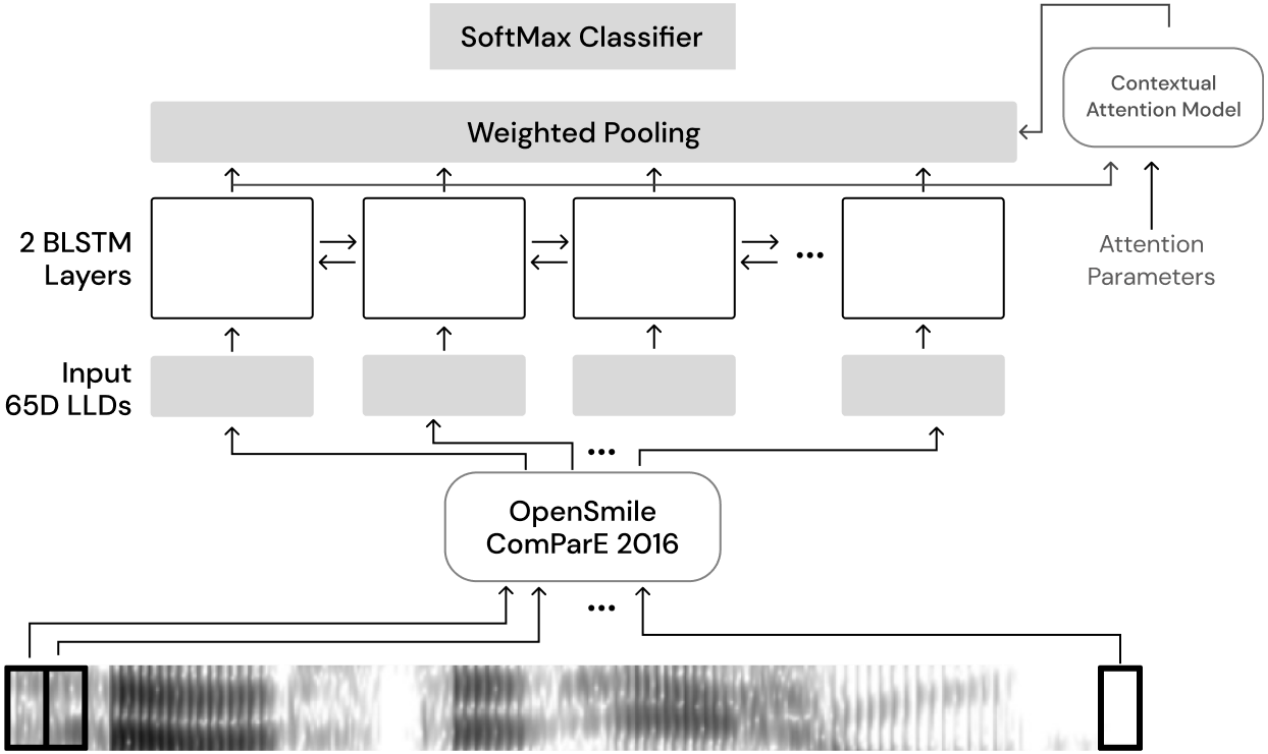
$$a_i = \frac{\exp(e_i)}{\sum_{j=1}^{N} \exp(e_j)}, \quad \text{where} \sum_{i=1}^{N} a_i = 1 \qquad z = \sum_{i=1}^{N} a_i h_i$$

Where h_i is the output from the i^{th} **BLSTM block**. v is the **attention vector** to be learned alongside the weight matrix W_h and bias vector b_h.

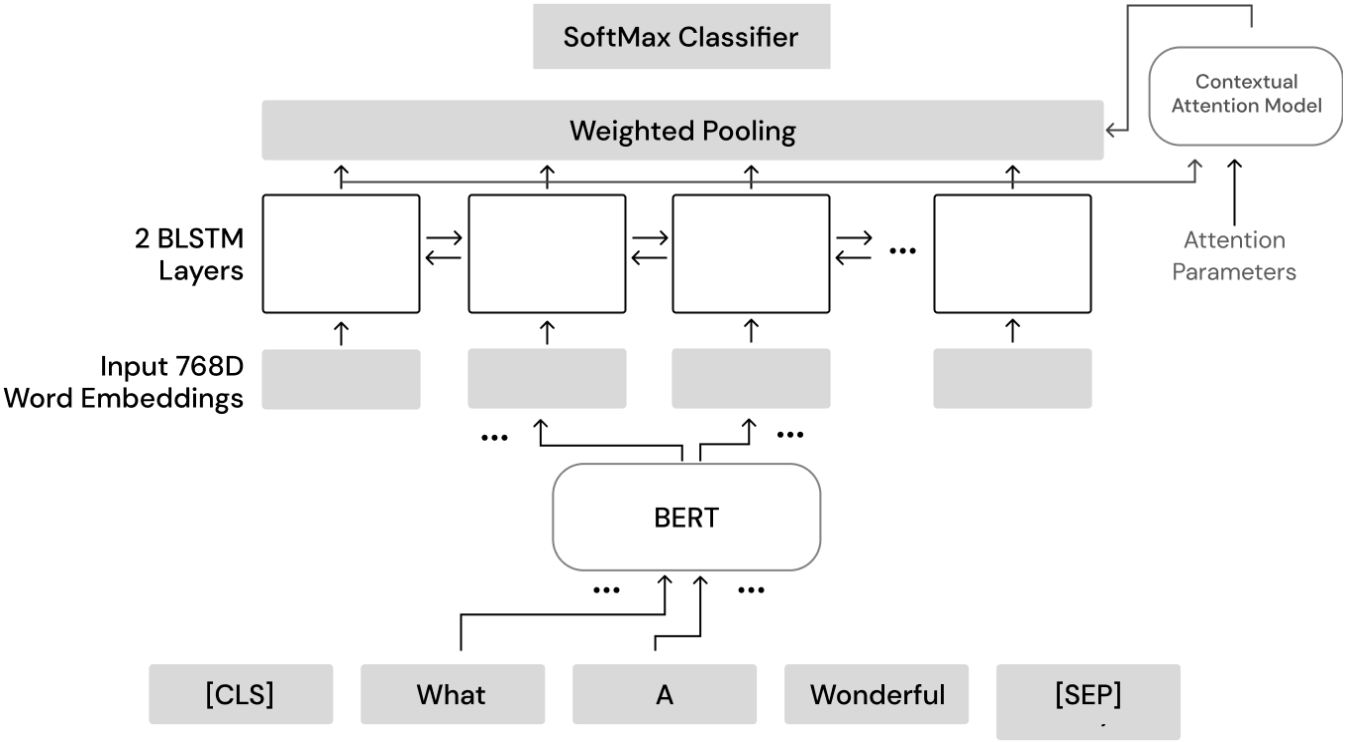


Ref. S. Mirsamadi et al. (2017), Bahdanau et al. (2014).

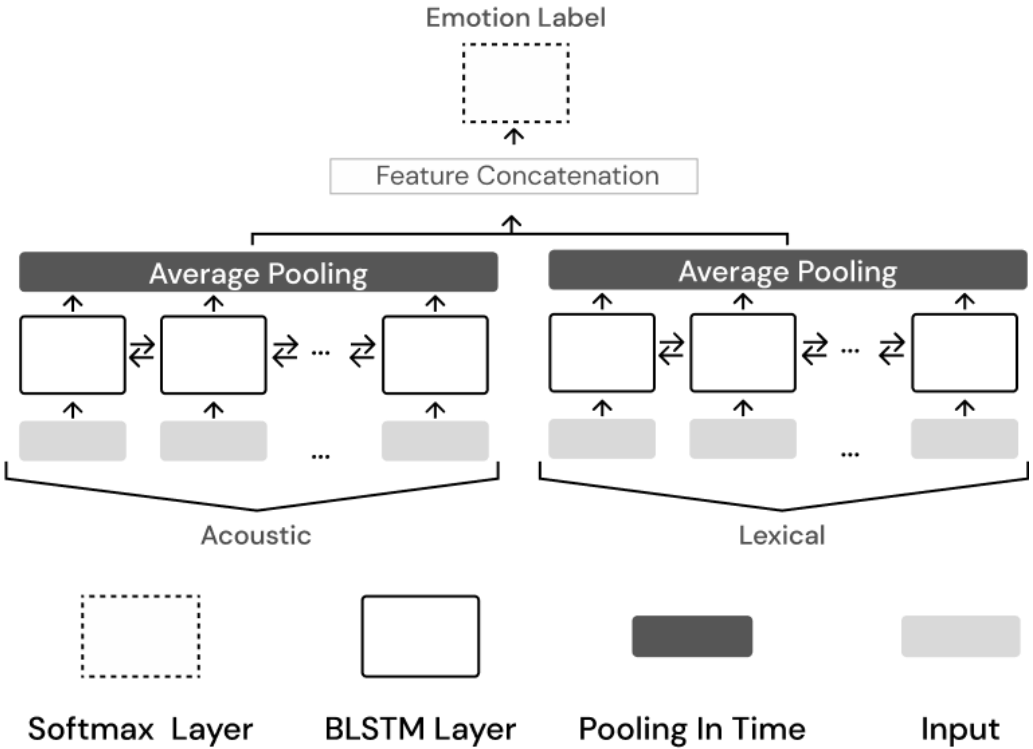
Architecture: Acoustic Model with Attention



Architecture: Lexical Model with Attention



Architecture: Multimodal Baseline



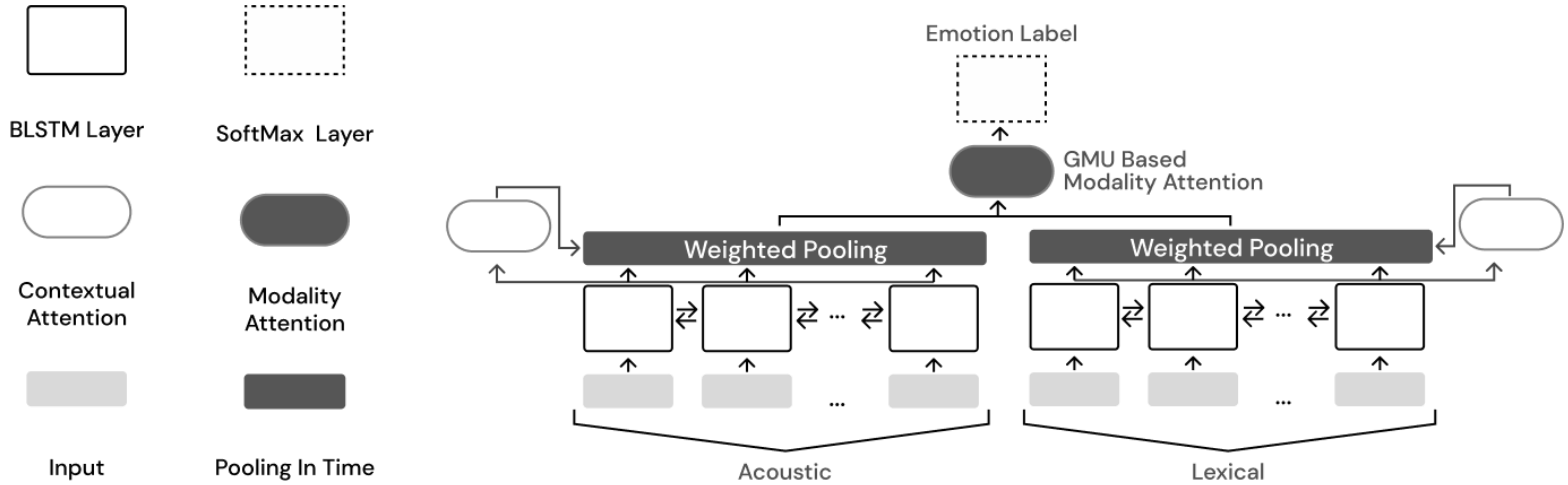
Ref. Gustavo et al. (2019)

Architecture: Proposed Multimodal Multi Focus Model

GMU: Modality-based attention Arevalo et al. (2017).

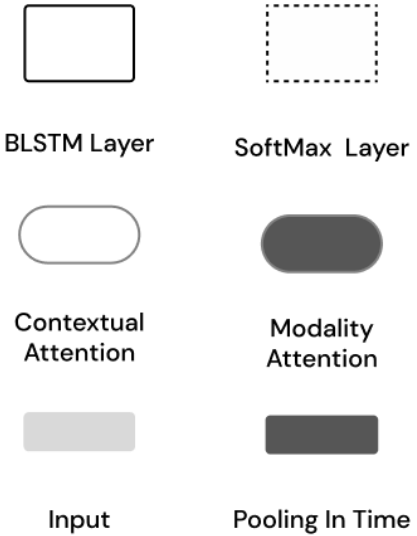
$$h_a = \tanh(\mathrm{W}_a x_a + b_a) \qquad z = \sigma(\mathrm{W}_z[x_a, x_l] + b_z)$$
$$h_l = \tanh(\mathrm{W}_l x_l + b_l) \qquad h = z * h_a + (1 - z) * h_l$$

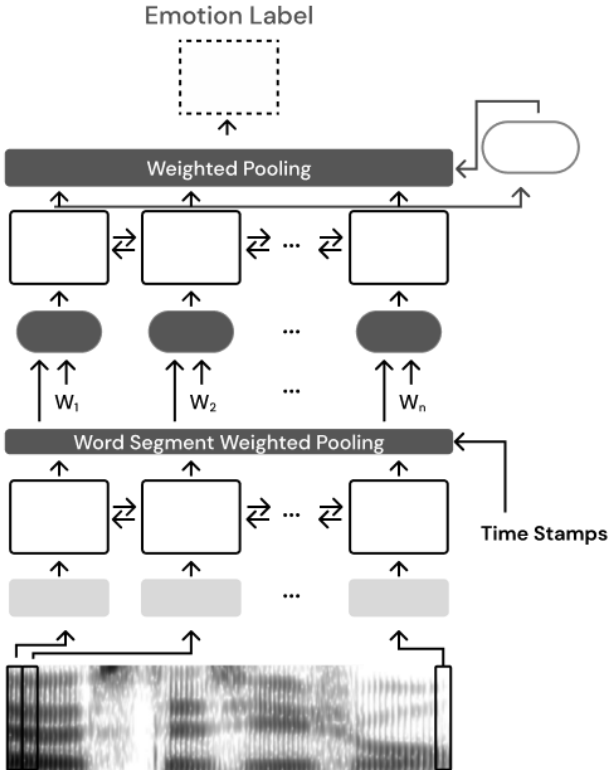
Here x_a and x_l are the input acoustic and lexical vectors. h_a and h_l are called the hidden acoustic and lexical vectors, respectively. z is used to get the **element-wise summation** weights for the two modalities.



Arevalo et al. (2017)

Architecture: Novel Hierarchical Attention Model (future work)





Fusing modalities earlier in time, such as at the word level, can potentially yield better results because it allows the model to capture fine-grained, **complementary information from both modalities at a more detailed level**. Ref. Gustavo et al. (2019)

Propose : **word segment pooling**

This approach can potentially improve the performance of the emotion recognition model by **preserving the contextual information present in the utterances**

Ref. Gustavo et al. (2019)

Results: Acoustic Only

Test Results:

Model	Avg Loss	Weighted Accuracy	Unweighted Accuracy	Angry Acc.	Happy/Excited Acc.	Neutral Acc.	Sad Acc.
Last Block	1.1296	49.82%	50.03%	38.79%	38.44%	57.91%	64.97%
Avg. Pool	1.0815	54.98%	54.45%	37.38%	49.38%	63.00%	68.02%
Attention Weighted Pooling	0.959	60.05%	60.89%	61.21%	46.88%	65.42%	70.05%

Observation: model **benefits significantly** from the **contextual attention** module .

*Speaker dependent results

Analysis: Acoustic Only

- Acoustic model performance was worst when using the last block output for classification.
- Average pooling improved results, but attention-based weighted pooling performed even better.
- **Pooling methods** outperformed the last block method due to their **ability to capture information from the entire utterance**, especially in **audio inputs with large sequence sizes** and numerous frames. **(avg for our data ~500 frames per sample)**
- **Attention-based** pooling **identified emotionally salient regions** in the RNN output and assigned them higher weights during pooling, resulting in better performance.
- Ex. unlike **average pooling**, which **treats silent and speech regions equally**, **attention-based** pooling **prioritizes speech regions, as demonstrated in the results**. Ref. S. Mirsamadi et al. (2017)

Ref. S. Mirsamadi et al. (2017)

Results: Lexical Only

Test Results:

Model	Avg Loss	Weighted Accuracy	Unweighted Accuracy	Angry Acc.	Happy/Excited Acc.	Neutral Acc.	Sad Acc.
Last Block	0.9051	63.04%	61.50%	56.54%	68.12%	67.02%	54.31%
CLS Block	0.9369	63.13%	63.18%	62.62%	65.00%	61.66%	63.45%
Avg. Pool	1.0551	62.68%	60.56%	52.34%	70.31%	67.83%	51.78%
Attention Weighted Pooling	0.9641	63.13%	64.83%	73.36%	55.62%	59.79%	70.56%

Observation: model performance improves with the **contextual attention** module but **not significantly.**

*Speaker dependent results

Analysis: Lexical Only

- All three methods (last block output, pooling, and attention-based) **performed similarly** for the lexical model. **Attention** methods did **not show significant performance improvements**.
- Last block method performed slightly better than average pooling method and similarly to the attention-based pooling method due to **the short length of input sequences (words in an utterance).**
- The success of the last block method is relatable to **BERT's architecture which is good at capturing the global context in the input sequence. (High-level Features compared to LLDs through OpenSmile for acoustic)**
- Using the **CLS token in BERT also yielded good results** for the lexical model.

Results: Multimodal Models

MM-B (Baseline Multimodal) Results

Test Performance

Avg Loss	Weighted Accuracy	Unweighted Accuracy	Anger Acc.	Happy/Excited Acc.	Neutral Acc.	Sad Acc.
0.6544	75.72%	75.99%	77.33%	75.14%	74.47%	77.00%

MMMLA (multimodal multi level attention) Model Results

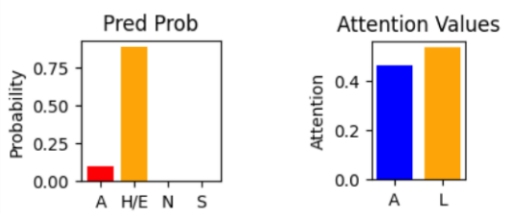
Test Performance

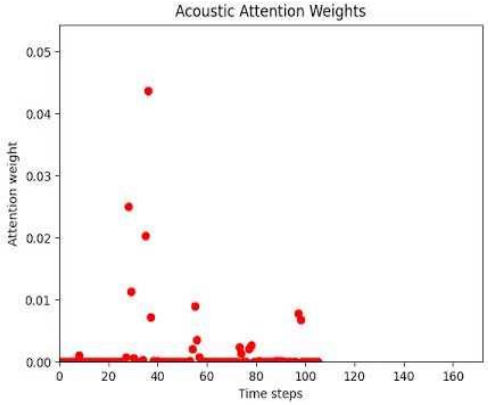
Avg Loss	Weighted Accuracy	Unweighted Accuracy	Angry Acc.	Happy/Excited Acc.	Neutral Acc.	Sad Acc.
0.5862	77.72%	78.31%	83.11%	75.14%	75.98%	79.00%

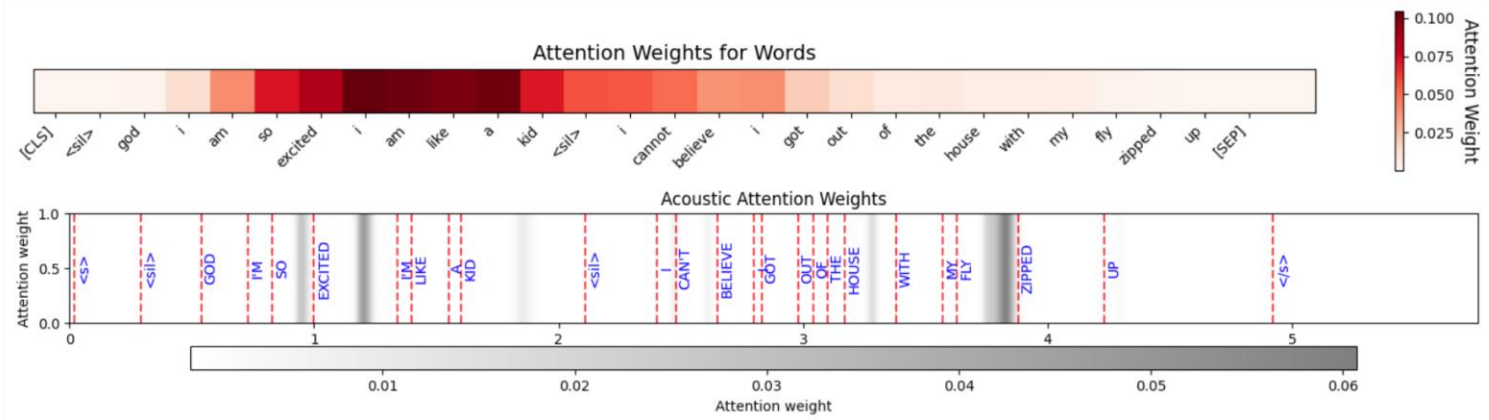
*Speaker dependent results

Results and Analysis: MM-MLA

Predicted Label: hap
True Label: hap
trans_words: <s> <sil> GOD I'M SO EXCITED I'M LIKE A KID <sil>
CAN'T BELIEVE I GOT OUT OF THE HOUSE WITH MY FLY ZIPPED UP </s>

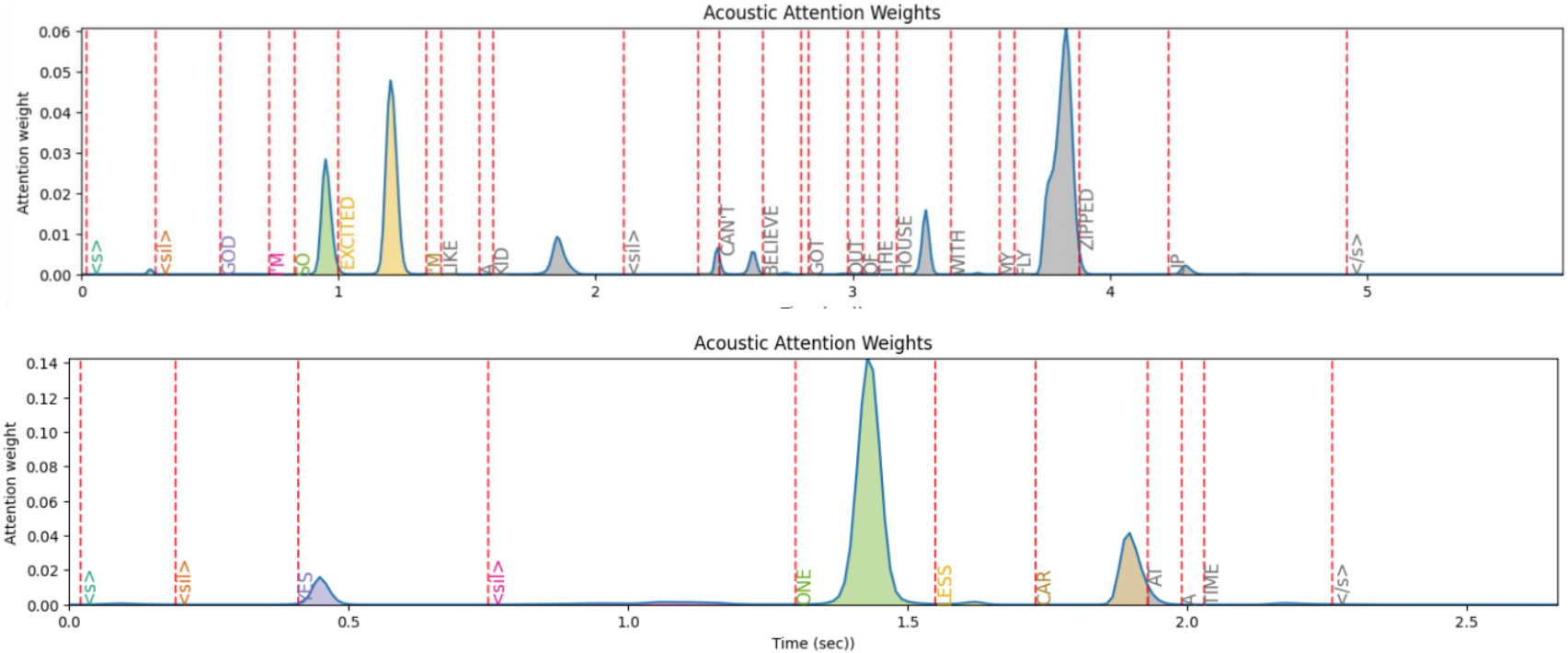






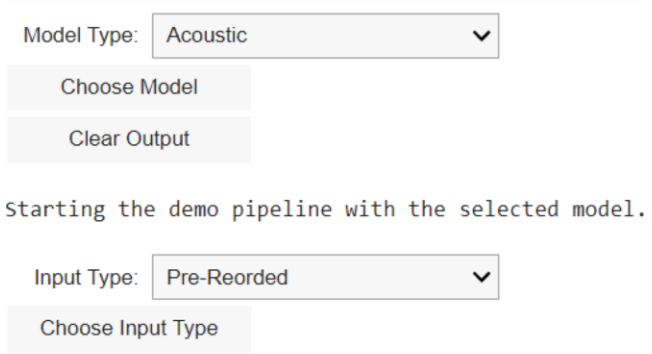
Results and Analysis: MM-MLA

model has **automatically learned to reject silent region** and likely region where the **emotional information is likely to be low**.

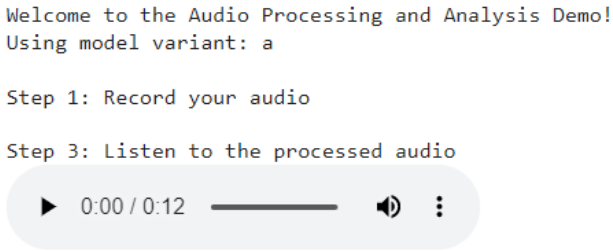


Demo

The interface is shown below and allows user to **choose** a **model** and **input type**:



Then **audio can be recorded and reviewed**:



Demo

Then the **transcript** is generated from the input using a **ASR module,** followed by the extraction of acoustic and lexical features. If for some reason the **ASR model is not able to generate transcript** we **default to the acoustic** model to get the results.

Finally, model is loaded to perform predictions and give predicted output with class probabilities.

User gets the option of trying again with a new input.

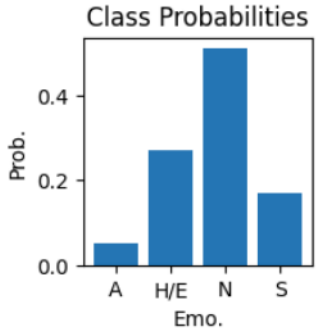
```
Step 4: Getting the transcript...
Transcript: <s> this yard this chair when you away from anything </s>

Step 5: Extracting acoustic features...
Acoustic features extracted, shape: torch.Size([1, 1232, 65])

Step 6: Extracting lexical features...
Lexical features extracted, shape: torch.Size([1, 11, 768])

Step 7: Loading the pre-trained model...

Step 8: Performing predictions...
Class probablities: [[0.04983579 0.26964837 0.5109408  0.16957502]]
prediction: neutral
```




```
Demo Complete!
```

Run again

Evaluation Scheme

1. Attractive and complex problem: 5
2. Clarity in task and in input-output description: 5
3. Dataset effort- collection, annotation: 10
4. Workflow, Architecture, technique: 10
5. Results and analysis: 10
6. Demo working: 20