A Multi-modal approach to Speech Emotion Recognition

Jiayu Zhou, Paras Sibal, Rohit Bernard, Sanjana M Mallikarjuna, Swapnil Mallick
University of Southern California

USC Viterbi School of Engineering

MOTIVATION

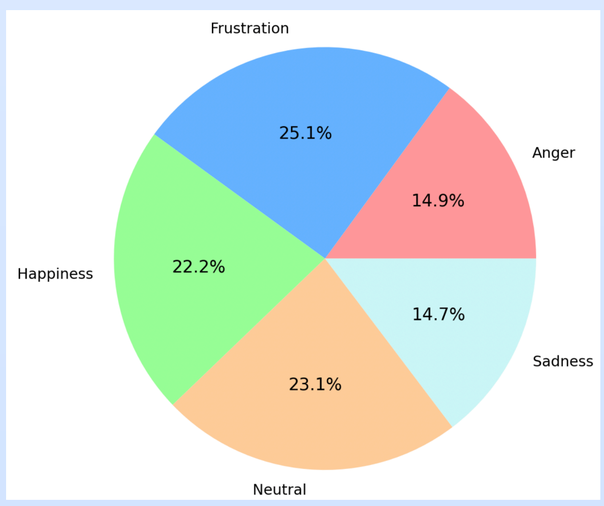
- Speech Emotion Recognition (SER) attempts to identify emotions from speech irrespective of the semantic content.
- Recognition of emotions in speech presents several challenges. Emotions are very subjective, and sometimes difficult for humans to identify appropriate emotions expressed in normal conversation.

Why Multi-modal?

- This project aims to design a multi-modal architecture on a labelled dataset instead of comparing deep learning models used by existing research.
- Currently, there are various models built separately for Speech to Emotion and Text to Emotion on unrelated datasets.
- We intend to explore the same dataset on a model that has transcripts as features along with audio data for the prediction of emotions. We want to explore the accuracy of such a model in comparison with the models we build separately for Speech and Text.

DATA COLLECTION

- Dataset: USC's Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) dataset.
- There are 10039 utterance files generated from ten actors.
- 5 emotion categories: **Anger, Frustration, Happiness, Neutral & Sadness.**



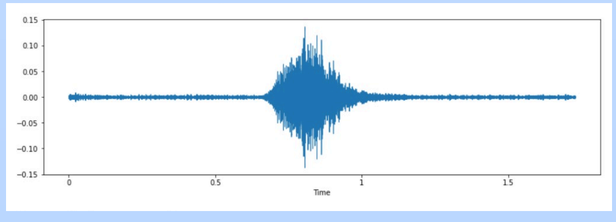
- Each utterance can have more than one emotion because it is possible that each annotator labels an utterance with more than one emotion. This shows that nature of human interaction from the hearing can lead to different meaning.
- For each utterance, we extract audio, transcription, percent of each emotion that can be used to describe that utterance based on annotator's label assignment.

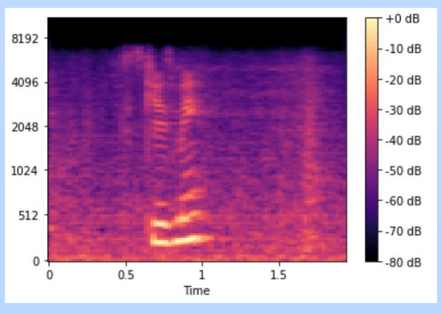
```
{'transcription': "I'm there in front of them. I'm telling them what
I'm good at.", 'emotions': "{'Frustration': 0.8, 'Anger': 0.2}"}
```

PREPROCESSING

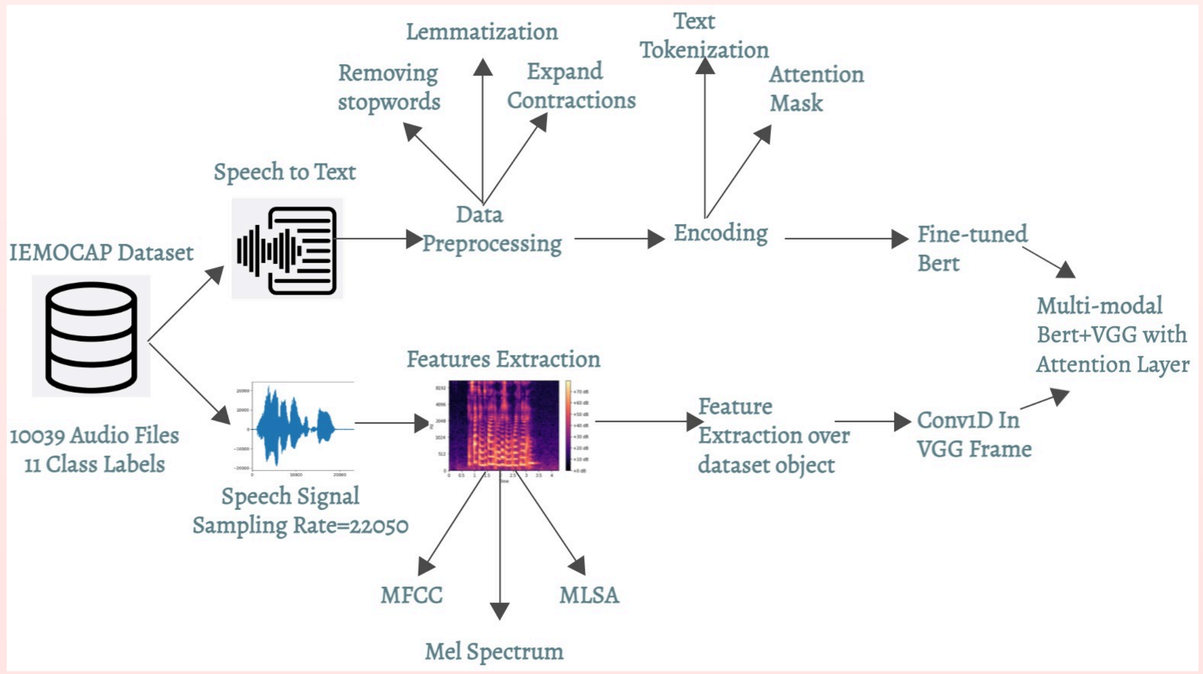
Preprocessing is a crucial phase to tune audio data and create feature vectors used as input to the model.

- Zero padding to shorter raw audio clips with non-uniform dimensions (varying lengths).
- Mel Spectrograms are converted to Mel scales to reduce perpetual differences. Hence making it easier to deal with human frequencies.
- Mel Spectrograms are compressed into short-time spectrums to extract only essential coefficients - MFCC, which only corresponds to human frequency ranges.
- Data Augmentation: Noise, Stretch, Pitch & Shift.





DESIGN



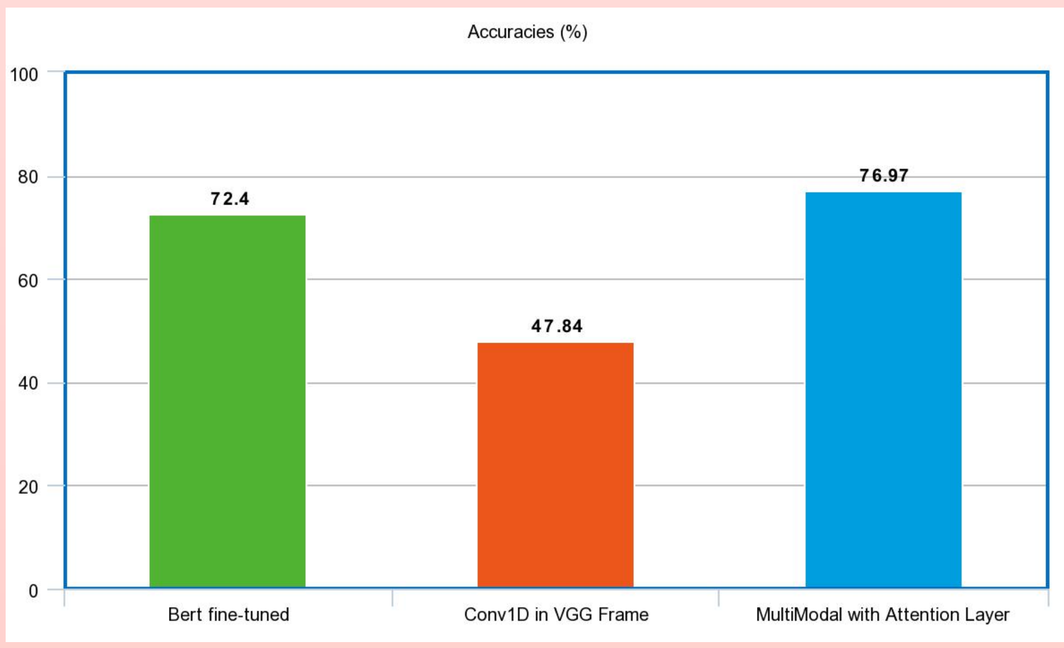
Built three models:

- Text to emotion - fine tuned BERT
- Speech to emotion - Conv1D in VGG frame
- Multimodal architecture - BERT + VGG with attention layers.

We evaluate the each model on an identical train-test split of the dataset.

RESULTS

- We have made a performance comparison between text-only to emotion vs audio-only to emotion vs multimodal where transcriptions of the audio are considered as a feature to audio data.
- As we can see from the preliminary results, the Multimodal model achieves only a slight improvement over a text-only based model, but gives a significant improvement over an audio-only based model.
- We also noticed, that audio based models perform better on single words, as compared to complete sentences.



Example Output:

- "Cool, perfect. Another Chicagoite, got to love it." ——>😄
- "This is ridiculous. I- I seriously, I don't understand why you think these automated systems are supposed to like work for anybody. They have never, ever work for me." ——> 😡

REFERENCES

- Gaurav Sahu. 2019. Multimodal speech emotion recognition and ambiguity resolution. CoRR, abs/1904.06022.
- Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung.2018. Multimodal speech emotion recognition using audio and text. CoRR, abs/1810.04635.