Group 7 Final Project Report
Project Title: Campaign Speech Emotion Recognition

Team Members:
Chris, Samiksha, Faisal, Meng-Fei

Date: 2024/12/11

1. Introduction

Speech is an important part of communication among human beings and is a way to express thoughts, feelings, and moods. Beyond just information being transferred, emotion plays an important role in speech communication and emotional speech is an effective way of conveying a message. Subtle changes in lexical or grammatical emphasis can change the meaning of a statement, with changing the words (1). This fact can be best demonstrated by the fact that a voice call is more informative than a text message.

Speech Emotion Recognition (SER) is a field of study which focuses on inferring human emotion from speech signals. SER is emerging as an important topic with applications in healthcare, machine-human interaction, education, intelligent assistance and many others (2). Many questions of SER are classification problems and can be addressed with supervised machine learning methods.

This project focuses on SER applied to political campaign speeches. By detecting emotions such as anger, happiness, fear, and neutrality from voice alone, the goal is to gain deeper insights into the emotional tenor of political messaging. Understanding these emotional cues can help decode communication strategies and the underlying sentiments intended to influence public opinion.

Each member contributed a unique approach:

Samiksha: A CNN-based model using Mel spectrograms.

Faisal: Initially Wav2Vec2-based embeddings with a fully connected classifier, later extended to using HuBERT embeddings for comparison and evaluating the model's generalization on external audio clips.

Chris: Audio Spectrogram Model and A multimodal fusion of ASR, SER, and textual embeddings.

Meng-Fei: A hybrid LSTM-CNN model using MFCC features, along with exploratory data analysis (EDA) , data visualization, and demonstration using Streamlit.

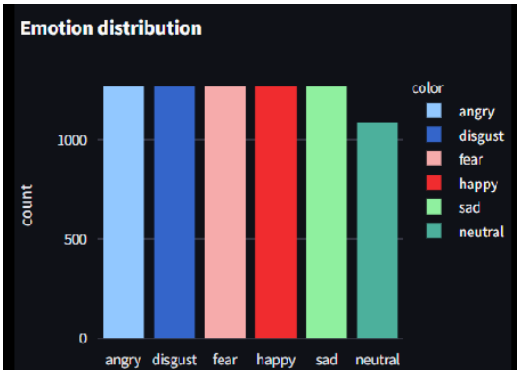
This report covers the dataset used, the methodologies applied, the experimental setup, the results obtained, and the final conclusions.

Report Structure:

- Section 2: Dataset Description
- Section 3: Deep Learning Networks and Training Algorithms
- Section 4: Experimental Setup
- Section 5: Results
- Section 6: Summary and Conclusions
- Section 7: References

2. Description of the Dataset

The CREMA-D dataset has been chosen for this project. It comprises 7,442 audio clips recorded by 91 actors, including 48 males and 43 females, aged between 20 and 74, representing diverse racial backgrounds. These clips capture one of six emotional states: Anger, Disgust, Fear, Happiness, Neutrality, and Sadness. CREMA-D's diversity in speaker demographics and emotional expressions makes it an excellent resource for training and evaluating Speech Emotion Recognition (SER) models. The distribution of the emotions reveals that the dataset is well-balanced. The sound waveforms suggest that the features of different emotions exhibit distinct patterns.



	Emotions	Count
0	angry	1,271
1	disgust	1,271
2	fear	1,271
3	happy	1,271
4	sad	1,271
5	neutral	1,087

Figure 1. Exploratory data analysis

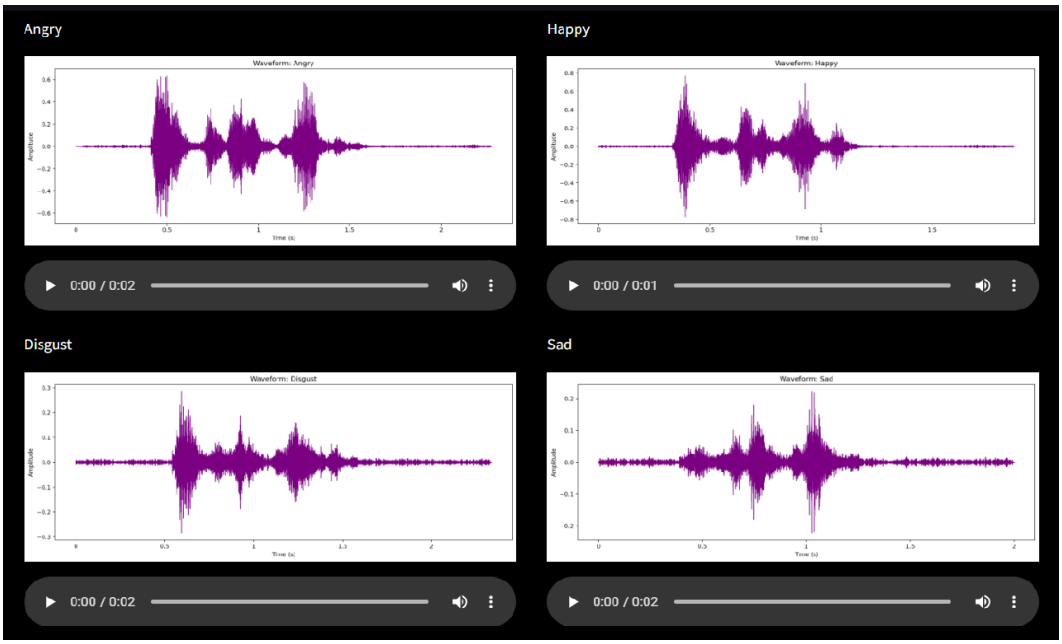


Figure 2. Sound waveform

3. Description of the Deep Learning Networks and Training Algorithms Approaches and Architectures:

1) **CNN with Mel Spectrograms:**
   A Convolutional Neural Network (CNN) model implemented in this project leverages Mel spectrograms of audio signals as input features. Mel spectrograms are a powerful representation of audio data, capturing both time and frequency variations, which are crucial for distinguishing emotional tones in speech. By leveraging 2D filters, the CNN effectively captures these variations and extracts complex spatial patterns from the spectrograms, enabling accurate classification of emotions. This approach highlights the capability of CNNs to identify subtle emotional cues in speech through their robust feature extraction mechanism.

   Model Architecture is shown here:

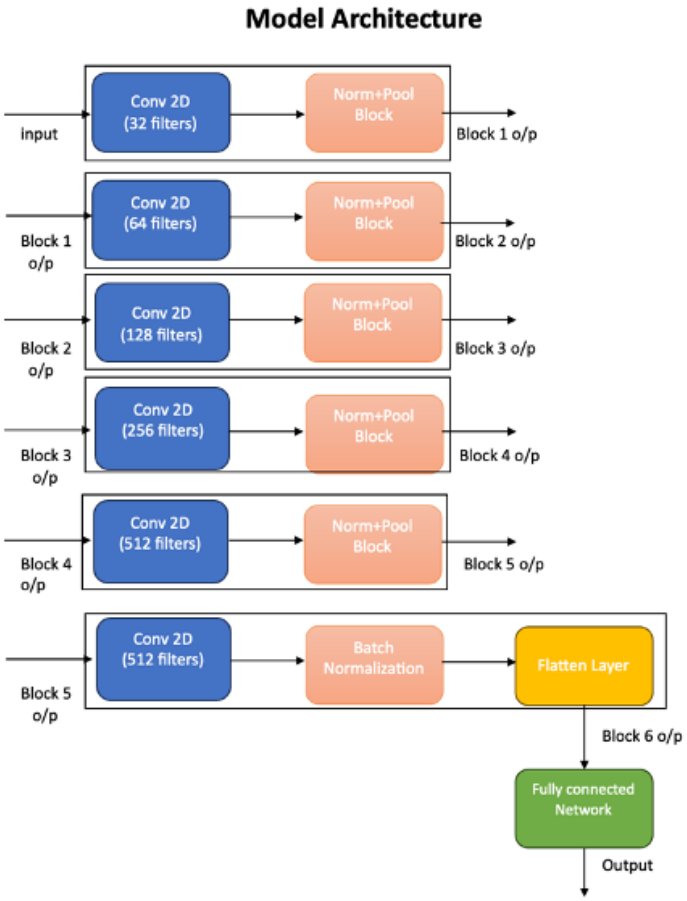


Figure 3. CNN model architecture

Key Features of the Model:

- 2D Convolutional Layers: Capture both spatial and temporal variations in the spectrograms, making the model adept at understanding complex emotional nuances in speech.
- Batch Normalization: Stabilizes and speeds up training by normalizing intermediate features.
- Pooling Layers: Ensure computational efficiency and invariance to small distortions in the input.

This architecture showcases how CNNs can leverage structured features from audio signals to identify emotions with precision. By progressively extracting patterns from simple to complex, the model effectively translates audio variations into emotional categories, demonstrating the power of convolutional networks in speech emotion recognition.

2) Integrate LSTM and CNN with MFCC Features:

A second approach used MFCC features, a classic representation in speech processing, fed into a hybrid LSTM-CNN model. This architecture combines CNN and LSTM layers to effectively process audio data. The model begins with two convolutional layers in CNN layers, designed to capture local patterns and short-term temporal features in audio signals. These are followed by LSTM layers, which learn temporal relationships and long-term dependencies in the sequential data. Finally, a fully connected layer translates the extracted features into accurate predictions.

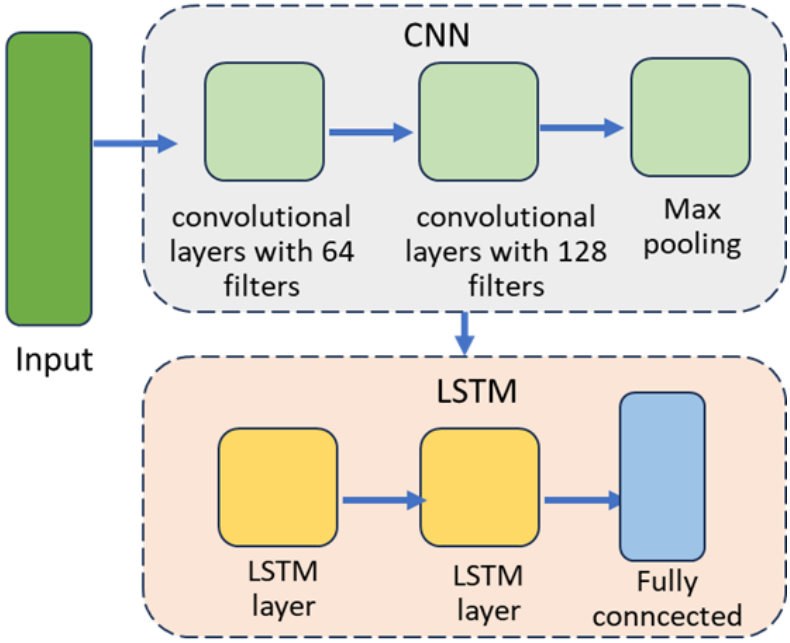


Figure 4. LSTM-CNN model architecture

3) Wav2Vec2 Embeddings with Dense Layers:

Another approach utilized Wav2Vec2, a pre-trained audio representation model. By extracting embeddings directly from raw audio and subsequently training dense layers for classification, this method capitalized on self-supervised features that often provide richer, more robust encodings of speech attributes.

HuBERT Embeddings:

Similar pipeline as Wav2Vec2, but using HuBERT for feature extraction. HuBERT's representations offer another form of self-supervised embeddings to compare against Wav2Vec2.

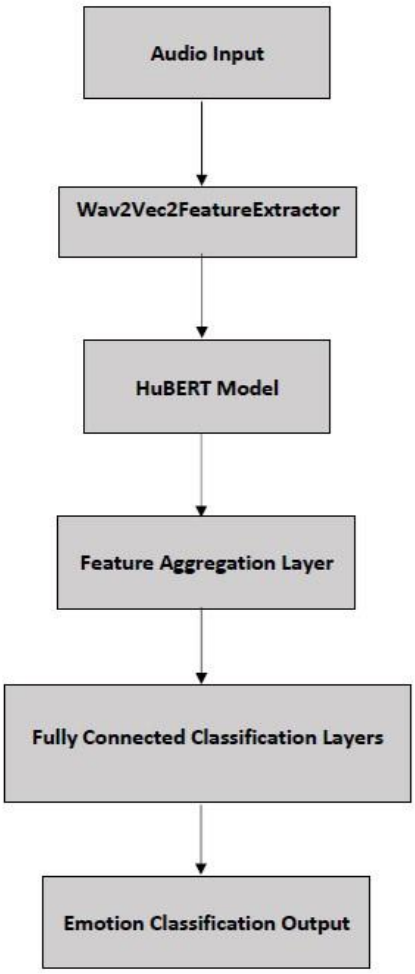


Figure 5. Pretrained model architecture

4) **Audio Spectrogram Transformer**

The Audio Spectrogram Transformer (AST) is a purely attention-based system that is applied directly to an audio spectrogram rather than the audio features (3). The AST takes an approach of transferring knowledge from the Vision Transformer model (4). The AST model takes audio input and converts it to a spectrogram image. The model interprets spectrogram images as a series of patches which contain information about time and frequency from the input. The attention mechanism of the model allows it to look at relationships between any parts of the audio, no matter the distance. The model architecture is shown here:

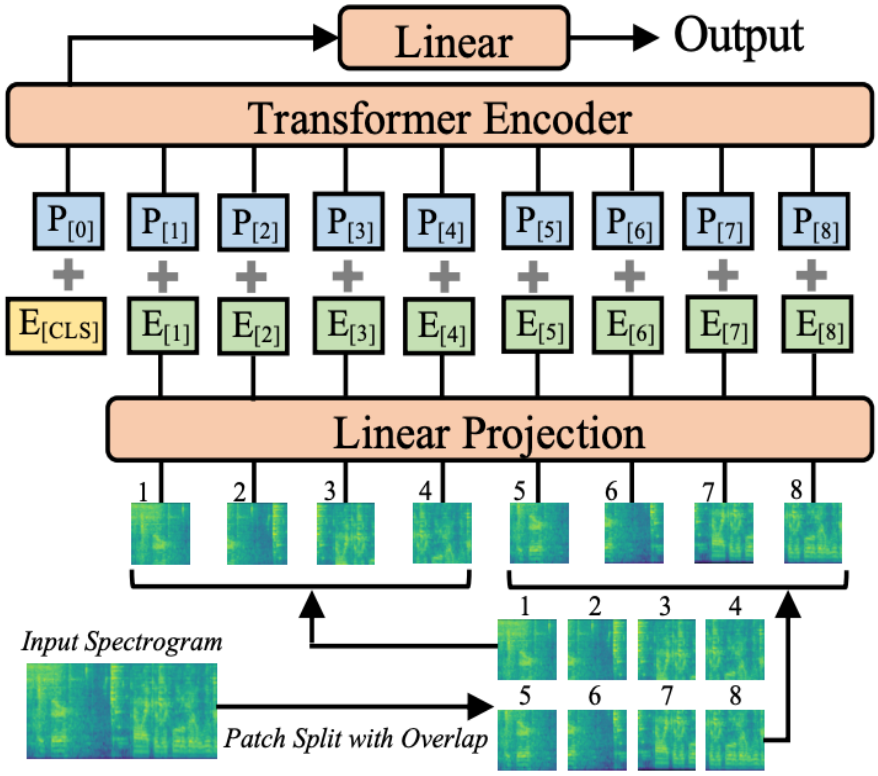


Figure 6. AST model architecture

5) Multimodal Fusion (Audio + Text):

This model incorporates several pretrained models to generate three modalities, SER, ASR, and text features. The embeddings are input to a multi-head attention layer. The multi-head attention layer aligns and integrates the three modalities and outputs a set of fused feature vectors that contains information from all modalities. The fully connected classifier layer produces the final output classification prediction.

Training Procedures:

All models used supervised learning with categorical cross-entropy loss. Adam or AdamW optimizers were used, and early stopping based on validation metrics helped prevent overfitting. Dropout layers and regularization techniques ensured that models generalized beyond the training data. Hyperparameters were tuned to balance training stability and performance.

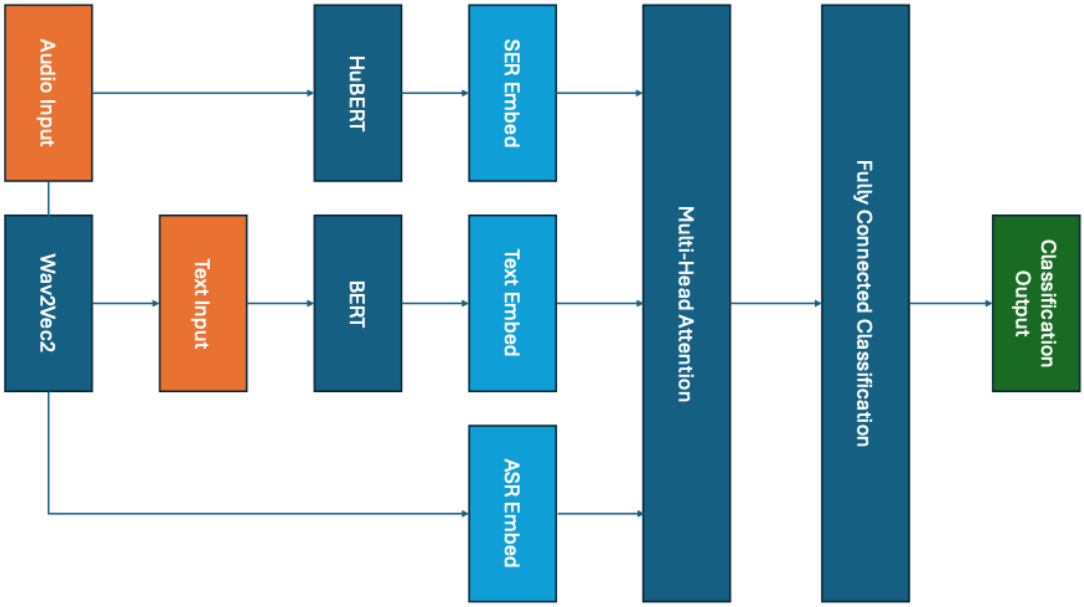


Figure 7. Multimodal model architecture

4. Experimental Setup

The approach to determine the best model was to compare the accuracy of the above models based on the CREMA-D dataset.

Data Preprocessing:

- Audio was resampled at 16 kHz for consistency.
- Mel spectrograms or MFCC features were extracted where required.
- The Wav2Vec2-based approach processed raw audio directly.
- For multimodal fusion, transcripts were generated via ASR, and text was tokenized using a BERT tokenizer.
- Stratified train/validation/test splits maintained balanced emotion distributions.
- Data augmentation (e.g., pitch shifts, added noise) was considered to enhance model robustness.

Frameworks and Tools:

- TensorFlow/Keras for CNN and dense-based models.
- PyTorch for LSTM and transformer-based architectures.

- Visualization tools (matplotlib, seaborn) and a Streamlit dashboard supported EDA, interpretability, and result visualization.

Performance Metrics:

- Accuracy served as the primary performance metric.
- F1-score and Cohen's Kappa were also considered to account for class imbalances and chance-level performance.

5. Results

CNN with Mel Spectrograms:

The CNN model achieved a test accuracy of 57.89%, with training accuracy steadily improving to approximately 97% over 30 epochs. However, the validation accuracy fluctuated, peaking at 60.37%, while the validation loss remained significantly higher than the training loss, indicating potential overfitting. Despite these challenges, the model demonstrated moderate effectiveness in distinguishing emotions, reflected by a macro F1-score of 0.56. The results for the custom CNN model are shown here:

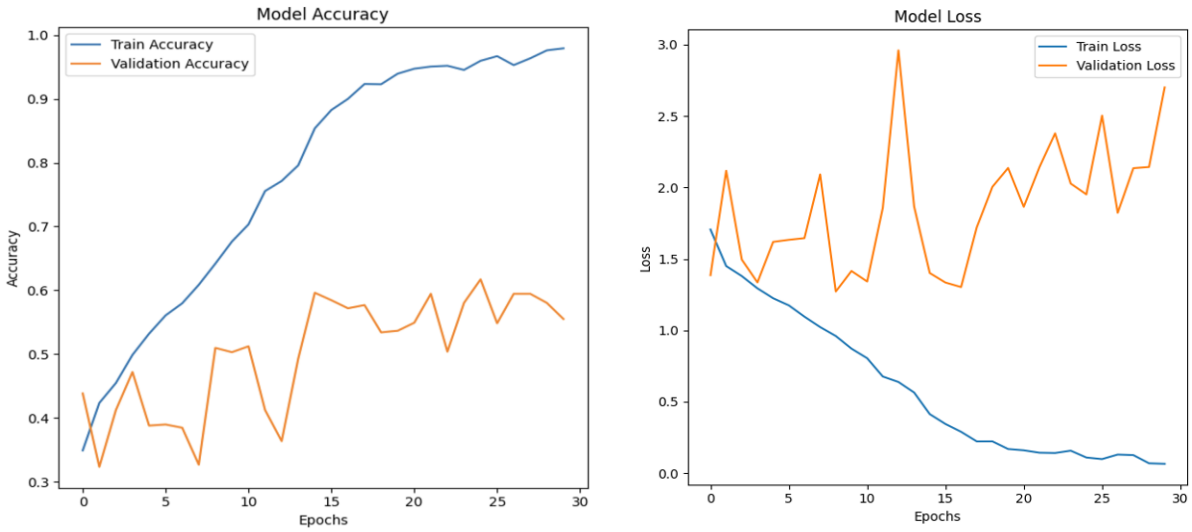


Figure 8. CNN accuracy

Table 1. CNN classification report

	precision	recall	f1-score	support
disgust	0.6334841628959276	0.5511811023622047	0.5894736842105263	254.0
happy	0.44543429844098	0.7843137254901961	0.5681818181818182	255.0
sad	0.5543859649122806	0.6220472440944882	0.5862708719851577	254.0
fear	0.6601941747572816	0.2677165354330709	0.38095238095238093	254.0
angry	0.7513513513513513	0.547244094488189	0.6332574031890661	254.0
neutral	0.6097560975609756	0.6880733944954128	0.646551724137931	218.0
accuracy	0.5742108797850907	0.5742108797850907	0.5742108797850907	0.5742108797850907
macro avg	0.6091010083197995	0.5767626827272603	0.5674479804428133	1489.0
weighted avg	0.6089752528311755	0.5742108797850907	0.56553595832375	1489.0

LSTM-CNN with MFCCs:

The hybrid LSTM-CNN model demonstrates only a modest improvement of 1–2% compared to the standalone LSTM model, achieving a maximum validation accuracy of 44.9%. This suggests that the model's capability for speech-emotion recognition is limited. One possible reason for the suboptimal performance could be the dataset's short audio clips, each lasting only two seconds, which may not provide enough data to effectively capture meaningful patterns for the LSTM component. In the future, incorporating audio datasets with longer durations or combining such datasets with the current ones could enhance the model's performance when using this hybrid approach.

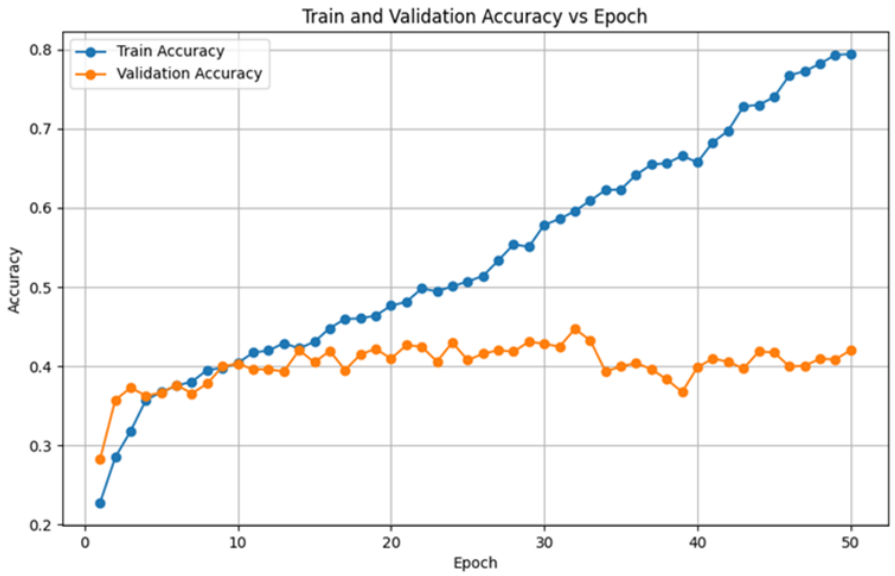


Figure 9. LSTM-CNN accuracy

Hubert Embeddings + Dense Layers:

Leveraging self-supervised audio embeddings improved accuracy to around 73%. The richer feature representations enhanced the model's ability to discriminate between closely related emotional categories.

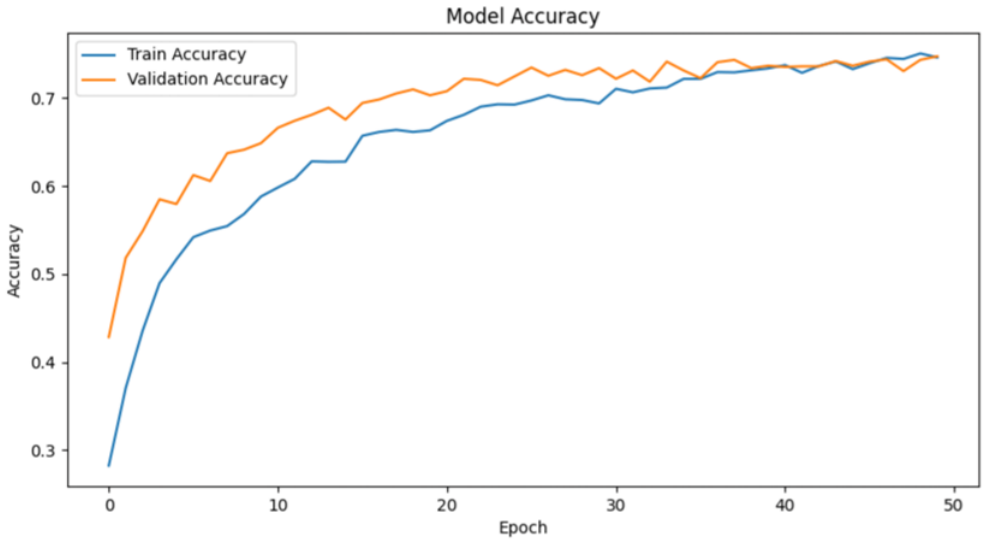


Figure 10. Pretrained model accuracy

Audio Spectrogram Model:

The AST model achieved a highest accuracy of 53.7%. This model incorporated audio augmentation to improve generalizability and robustness of the model. The model's use of spectrogram and transformer encoder with self attention allows the model to identify relationships in the audio no matter how far apart they are. The results of the AST model are shown here:

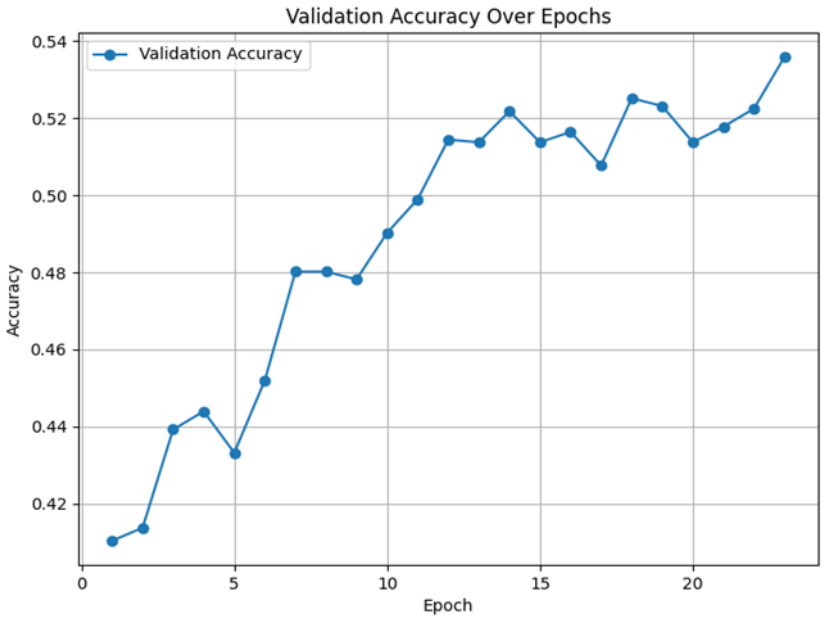


Figure 11. AST accuracy

Multimodal Fusion:

Preliminary results indicated that incorporating textual information via attention-based fusion could surpass 55.8% accuracy. The multi-head attention layer aligns and integrates the three modalities of ASR, SER, and texts then outputs a set of fused feature vectors that contains information from all modalities. The fully connected classifier layer produces the final output classification prediction. The results of this model are shown:

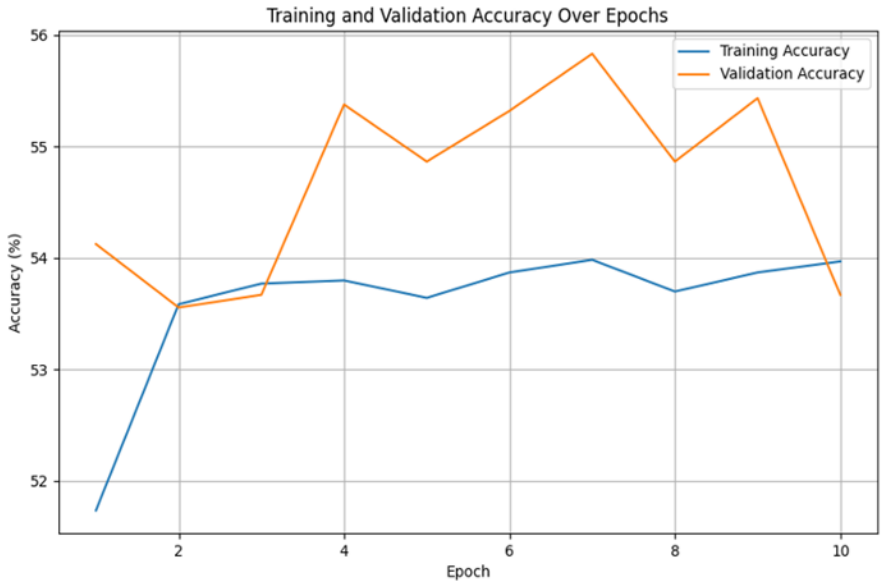


Figure 12. Multimodal accuracy

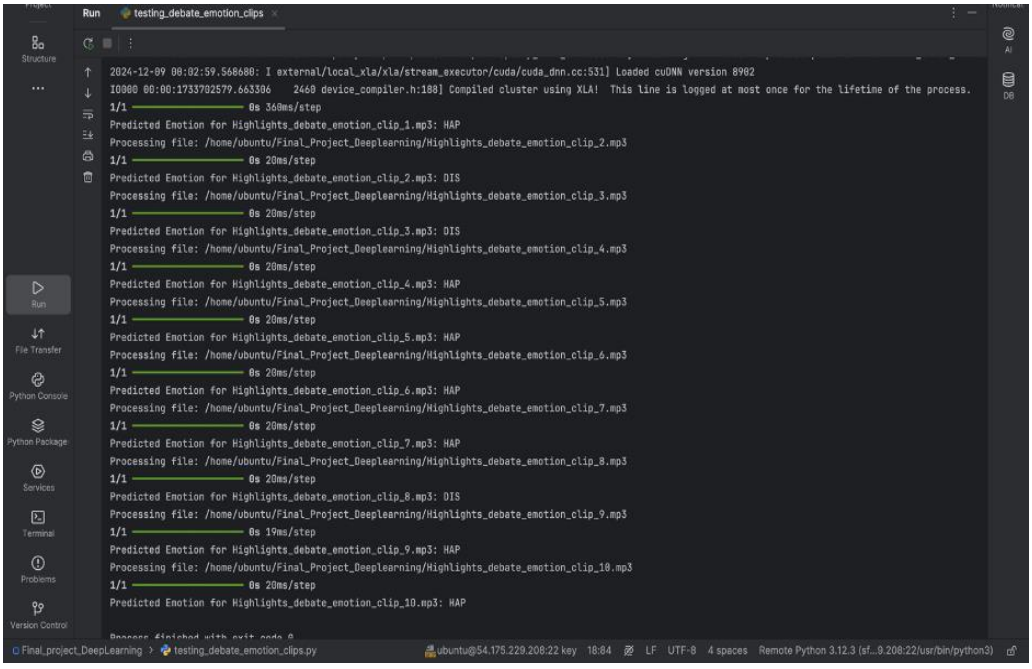
The pre-trained model outperforms the other four models by 10–20%, achieving the best performance among all the models mentioned above.

Table 2. Summary of Results

Models	LSTM-CNN	CNN	Multi	AST	Pre-trained
Validation Accuracy	40~45%	50~60%	55~60%	50~55%	70~75%

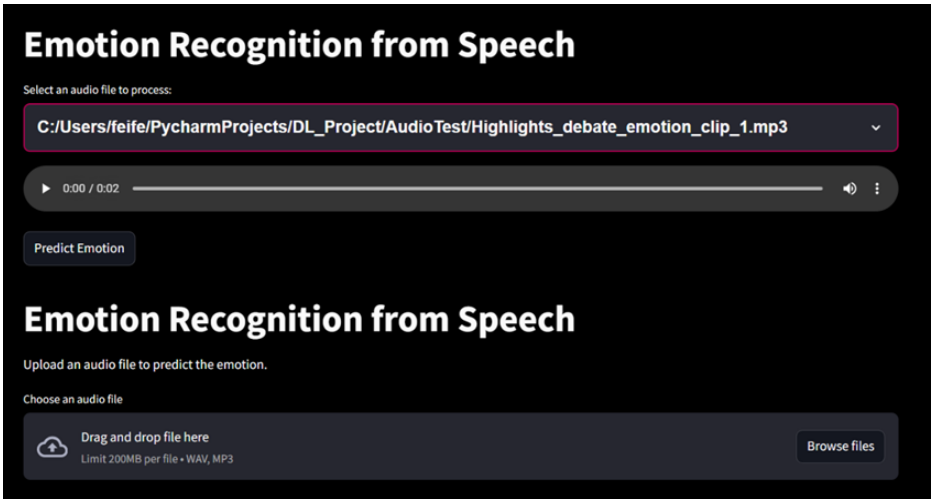
Testing on External Audio Clips:

Beyond CREMA-D, the final HuBERT-based model was tested on external audio clips—specifically, segments from political speeches. Since the model expects audio as input, properly formatted WAV files were used, or existing audio was converted to the correct sampling rate. The model produced predictions for these clips, demonstrating its flexibility. However, the predictions may not align with human interpretations of emotion in real-world political speech due to domain differences and the fact that the model never saw this kind of data during training. This highlighted a key limitation: while the model performed well on the curated CREMA-D dataset, it struggled to generalize to new domains. To improve real-world applicability, more domain-specific training data or domain adaptation techniques would be necessary.



Demonstration of speech emotion recognition(Streamlit):

We use Streamlit to create a dashboard that showcases our project research results. The dashboard displays elements such as datasets, exploratory data analysis (EDA), sound and waveform displays, and the architecture and performance of five models. Additionally, It imports the best-performing model, allowing users to explore preloaded political speech audio files as examples. Additionally, users can upload their own audio files to test the emotion recognition capabilities of the model.



6. Summary and Conclusions

Conclusion

This project investigated various approaches for Speech Emotion Recognition (SER), including a custom Convolutional Neural Network (CNN) leveraging Mel spectrograms, a combined LSTM-CNN architecture, HuBERT embeddings, the Audio Spectrogram Transformer (AST), and Multimodal Fusion (audio and text). Among these, pretrained models such as Wav2Vec2 and HuBERT achieved the highest accuracy, demonstrating the efficacy of self-supervised learning in capturing subtle emotional cues in speech. While traditional models like CNNs and LSTMs provided a solid baseline and offered insights into feature extraction, pretrained and multimodal approaches surpassed them in classification accuracy and robustness. The results emphasize the significance of leveraging advanced architectures and integrating multimodal data to enhance the performance of SER systems, paving the way for further exploration in real-world applications and more complex datasets.

Key Takeaways:

- Pre-trained embeddings (Wav2Vec2) generally improved SER performance, indicating the value of self-supervised learning for capturing subtle emotional cues.
- Multimodal fusion suggests that incorporating textual features can further enhance classification accuracy.
- Traditional features (MFCCs) and classic architectures (LSTM) remain informative for baselines and understanding model behavior.
- EDA, visualization tools, and interactive dashboards helped guide preprocessing decisions, model tuning, and interpretation of results.

Future Work:

- Apply these models to authentic campaign speech data for real-world validation.
- Explore advanced data augmentation, transfer learning, and domain adaptation techniques.
- Fine-tune multimodal architectures to fully exploit the synergy between linguistic and acoustic information.

7. References

1) Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, *21*(4), 1249.
2) George, S. M., & Ilyas, P. M. (2024). A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, *568*, 127015.
3) Gong, Y., Chung, Y. A., & Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.
4) Leelavathi, R., & ARUNA, V. (2021). Speech emotion recognition using LSTM. *International Research Journal of Engineering and Technology*.
5) Pan, Z., Luo, Z., Yang, J., & Li, H. (2020). Multi-modal attention for speech emotion recognition. *arXiv preprint arXiv:2009.04107*.
6) Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347-10357). PMLR.