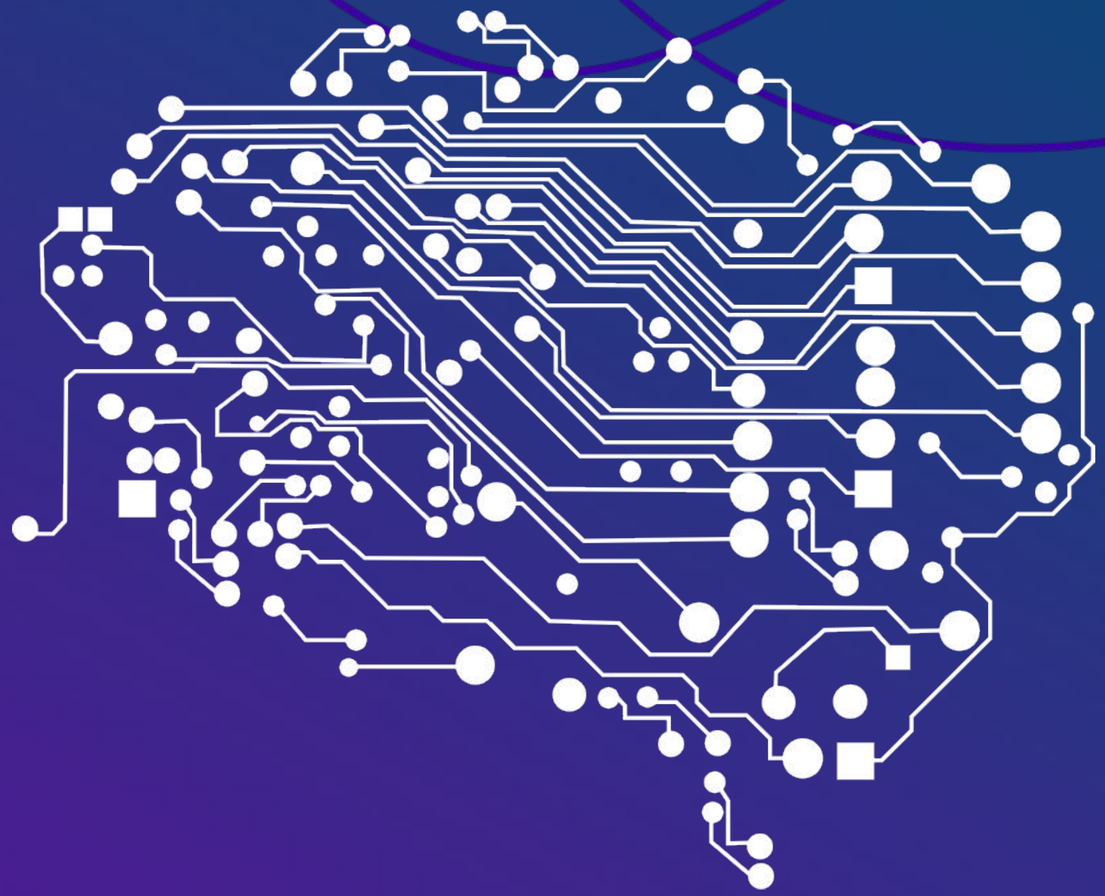
Speech Emotion Recognition (SER)

Vaibhav Kumar

Kanhaiya Kumar Sahu

Karan Deep Das

Chinmay Thakur



INTRODUCTION

Speech emotion recognition detects emotions from speech signals

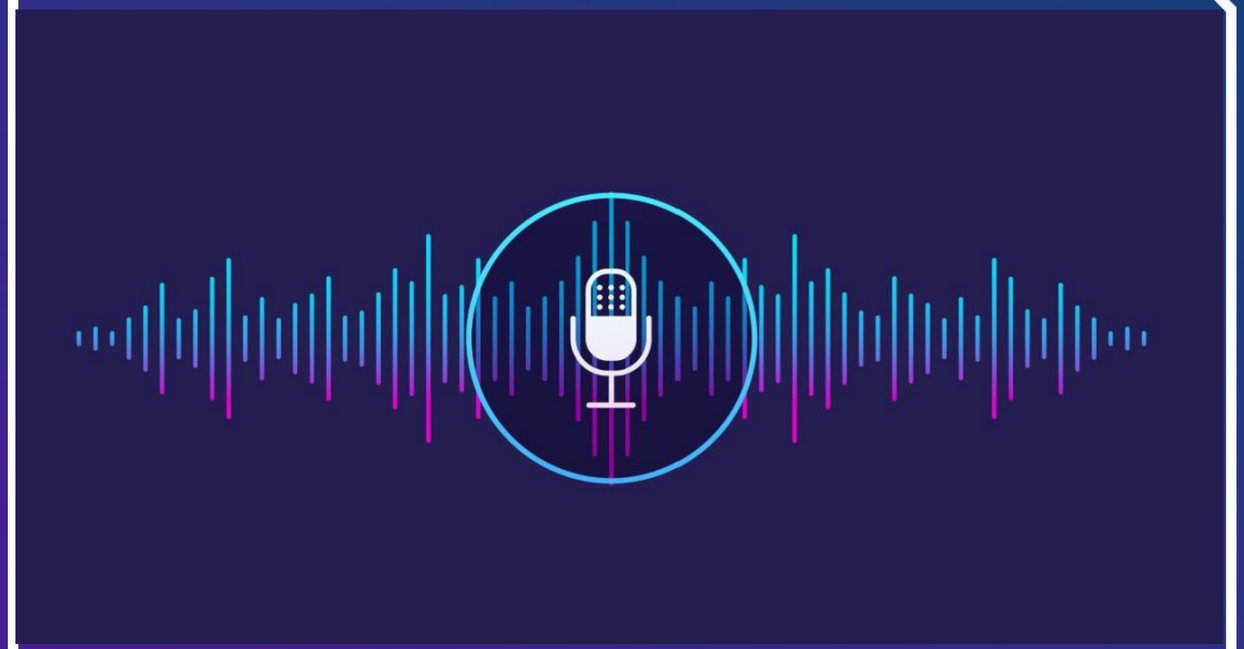
Used in mental-health monitoring, companion robots, enhanced call center analytics (customer support) and improving general human-computer interaction.

Problem formulation

- Core challenge : Accurate identification of human emotions from speech signals
- Key difficulties:
  - ❖ High Variability
  - ❖ Context Dependency
  - ❖ Acoustic Ambiguity & Noise

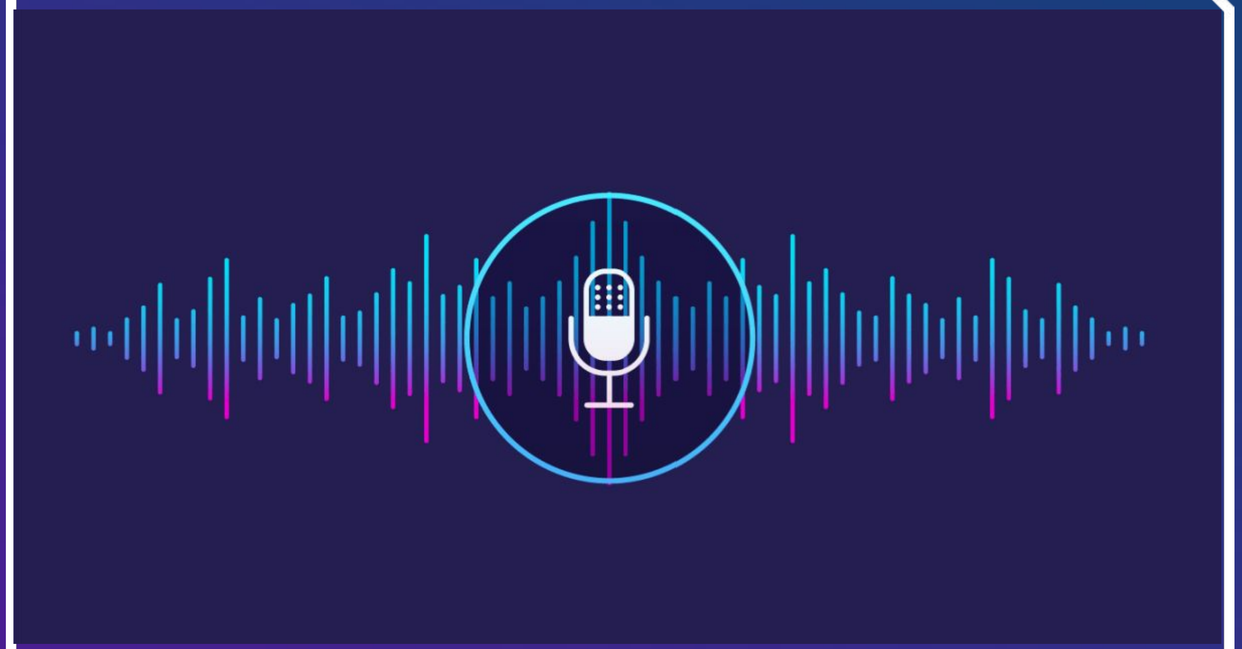
Problem formulation

- Key difficulties:
  - ❖ High Variability : Emotions manifest differently across speakers, accents and cultural backgrounds.



Problem formulation

- Key difficulties:
  - ❖ Context Dependency : The same utterance can convey different emotions depending on the conversational context



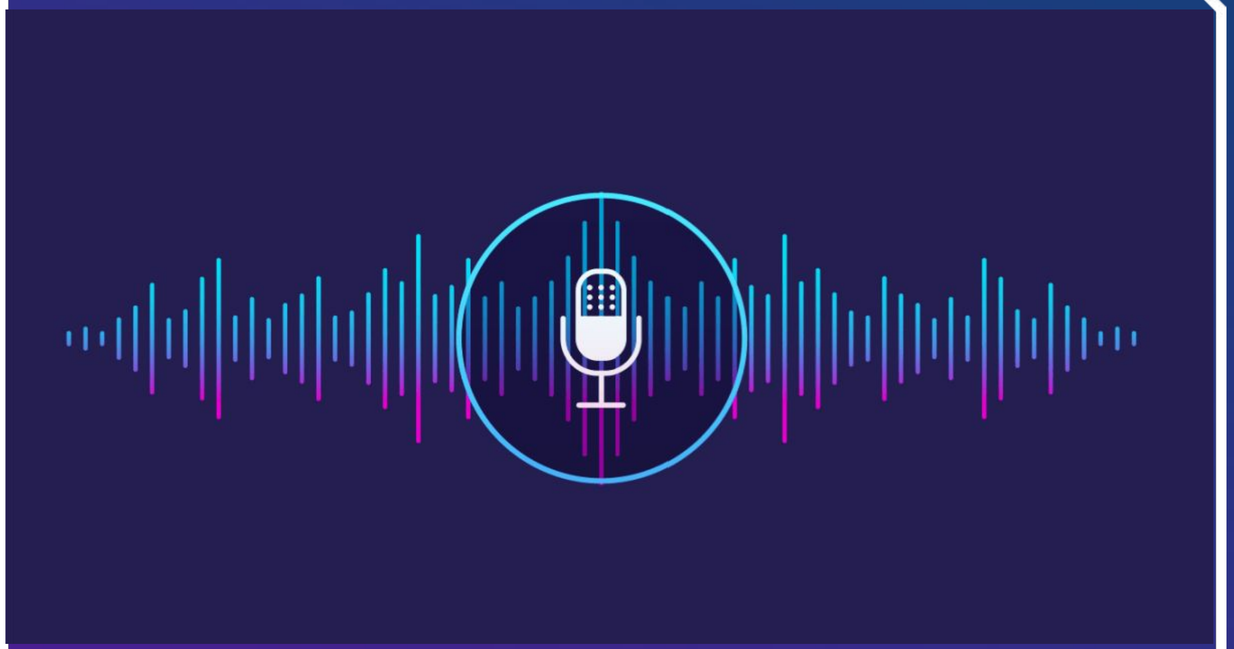
Problem formulation

- Key difficulties:
  - ❖ Acoustic Ambiguity & Noise : Emotional cues can be subtle, easily masked by background noise, or overlap between emotions

Problem formulation

- Our Goal :

  To develop a robust model that maximizes emotion classification accuracy by effectively modeling temporal dynamics and leveraging diverse acoustic features, while considering practical resource constraints.



Datasets & Preprocessing

To train and evaluate our Speech Emotion Recognition model, we utilized a combination of four publicly available datasets:

01

RAVDESS

- Ryerson Audio-Visual Database of Emotional Speech and Song
- 24 professional actors (12 male, 12 female).
- 8 emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprise.
- Speech-only files used.

Datasets & Preprocessing

To train and evaluate our Speech Emotion Recognition model, we utilized a combination of four publicly available datasets:

02

CREMA-D

- Crowd-sourced Emotional Multimodal Actors Dataset
- 91 actors (diverse ethnicities and ages).
- 6 target emotions: Anger, Disgust, Fear, Happy, Sad, Neutral (at varying intensities).

Datasets & Preprocessing

To train and evaluate our Speech Emotion Recognition model, we utilized a combination of four publicly available datasets:

03

TESS

- Toronto Emotional Speech Set
- 2 female actors.
- 7 emotions: Anger, Disgust, Fear, Happy, Pleasant Surprise, Sad, Neutral.

  (Note: "Pleasant Surprise" is often mapped to "Surprise")

Datasets & Preprocessing

To train and evaluate our Speech Emotion Recognition model, we utilized a combination of four publicly available datasets:

04

SAVEE

- Surrey Audio-Visual Expressed Emotion
- 4 male actors (native English speakers).
- 7 emotions: Anger, Disgust, Fear, Happy, Neutral, Sad, Surprise.

Data Unification & Filtering

- Audio files from all four datasets were combined into a single dataset.
- To create a consistent emotion set for our model, we focused on 6 primary emotions:
- Neutral, Happy, Sad, Angry, Fearful, Disgust.
- Emotions like "Calm" and "Surprise" (including "Pleasant Surprise") were filtered out from the combined dataset.

Tools and libraries

Programming language

python

Libraries

Numpy, Librosa , scikit-learn, seaborn , pandas , matplotlib, Transformers (from Hugging Face)

Deep learning framework

Keras

Feature Extraction Strategies

To capture relevant emotional information from speech, we employed a dual feature extraction approach

Mel-Frequency Cepstral Coefficients (MFCCs)

- A widely-used set of features representing the short-term power spectrum of sound, modeling human auditory perception.
- 39 MFCCs were extracted per frame (including delta and delta-delta coefficients).
- Sequences were padded/truncated to a fixed length of 162 frames to serve as input to our temporal modeling branch.

Feature Extraction Strategies

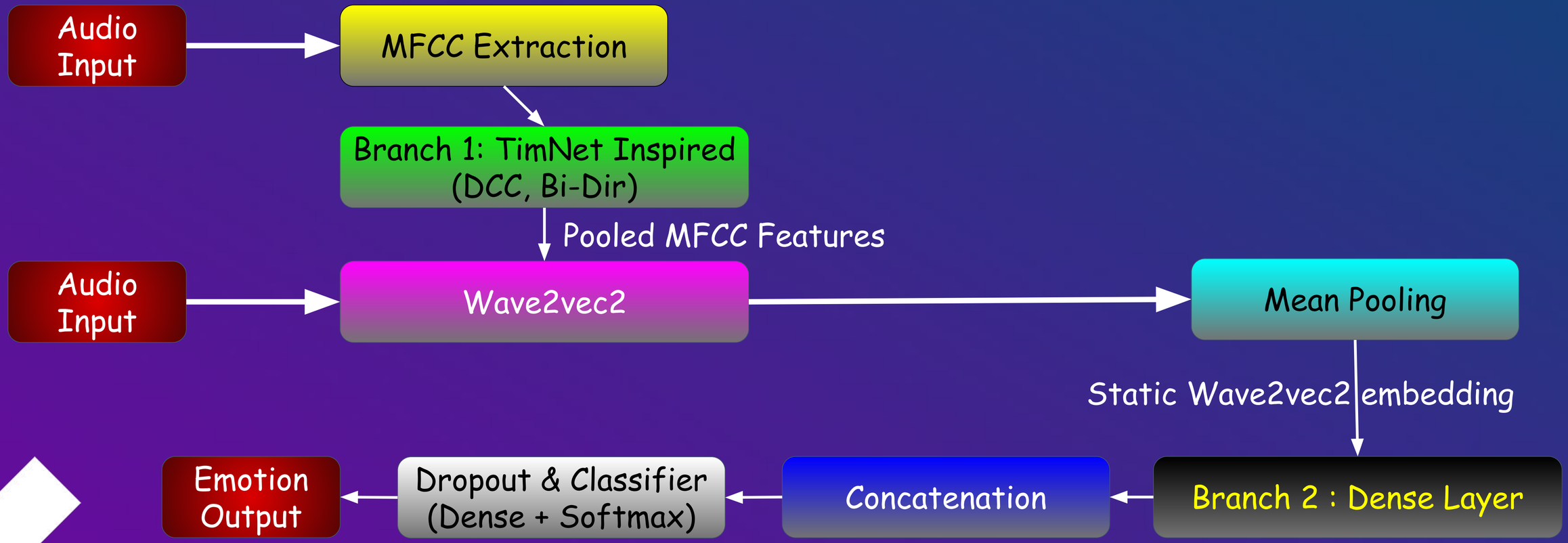
To capture relevant emotional information from speech, we employed a dual feature extraction approach

Pre-trained Wav2Vec2 Embeddings

- Leveraged a powerful, pre-trained Wav2Vec2 model (to generate rich contextualized speech representations.
- For each audio file, the Wav2Vec2 model's output (last hidden states) was mean-pooled across all time steps.
- This resulted in a single 768-dimensional embedding vector per utterance, capturing an overall representation of the audio.
- This static embedding was used as input to a separate branch in our ensemble model.

Proposed methodology - An Ensemble Approach

Our approach to Speech Emotion Recognition involves an ensemble model that integrates two distinct processing branches to leverage different strengths of feature representation



Proposed methodology - An Ensemble Approach

1. Data Preparation & Unification:

- Combined audio data from RAVDESS, CREMA-D, TESS, and SAVEE.
- Filtered to a consistent set of 6 emotions (Neutral, Happy, Sad, Angry, Fearful, Disgust) for focused training.

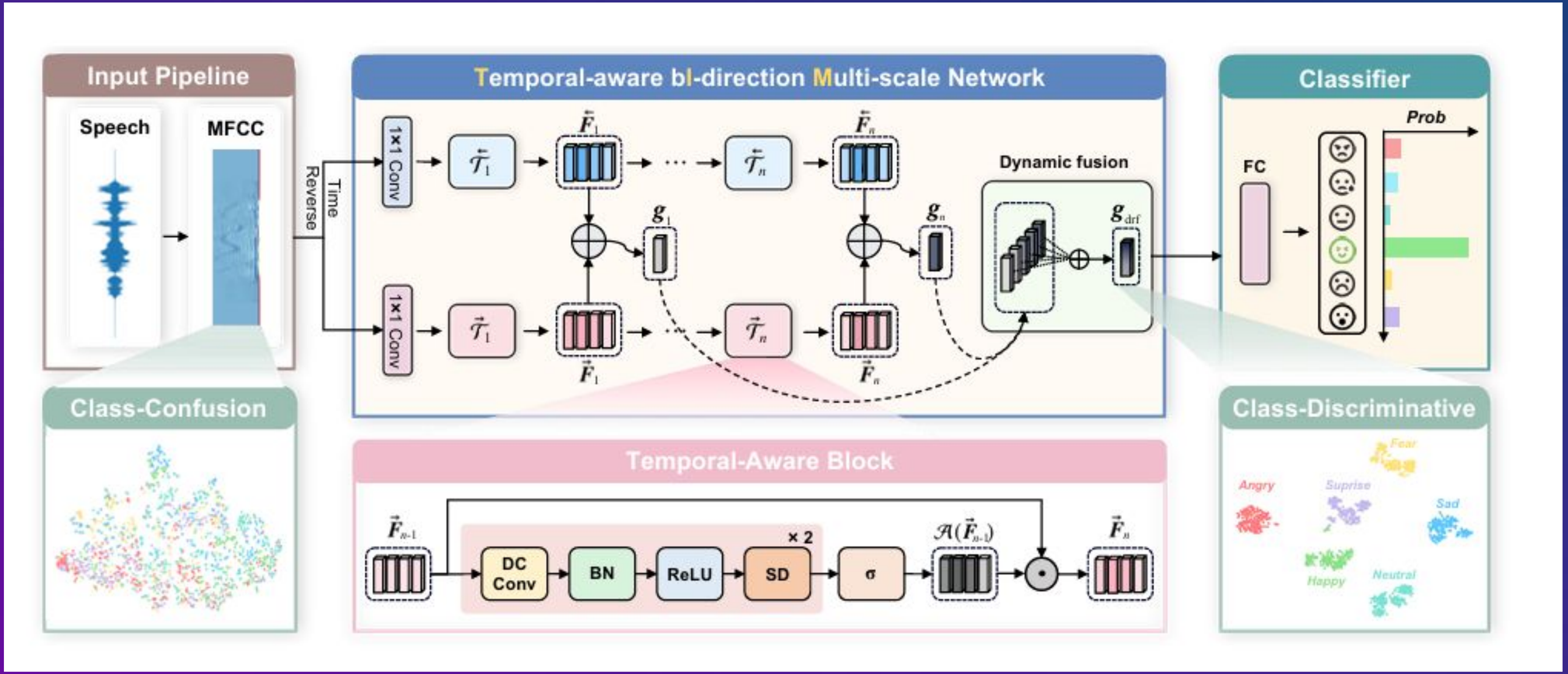
Proposed methodology - An Ensemble Approach

2. Ensemble Model Architecture:

- Branch 1: Temporal Modeling with MFCCs (TIM-Net Inspired):
    1. MFCC sequences (39 features, 162 frames) are processed.
    2. Utilizes a bi-directional structure with blocks inspired by TIM-Net's Temporal-Aware Blocks.
    3. Each block employs dilated causal 1D convolutions, batch normalization, ReLU, and spatial dropout to capture temporal patterns.
    4. Forward and backward pass outputs are added and then globally average pooled to produce a fixed-size temporal feature vector.
- Branch 2: Utterance-Level Embeddings with Wav2Vec2:
    1. Mean-pooled 768-dimensional Wav2Vec2 embeddings (derived from facebook/wav2vec2-base) provide a global representation of each utterance.
    2. These static embeddings are processed through a simple Dense (fully-connected) layer.
- Fusion and Classification:
    1. The feature vectors from both branches are concatenated.
    2. The combined feature vector is passed through a dropout layer and a final Dense layer with a softmax activation for multi-class emotion classification.

Inspiration: The TIM-NET framework and our Adaptation

Our temporal modeling branch for MFCCs was inspired by the TIM-Net framework (Ye et al., 2023), shown below, which excels at capturing multi-scale temporal emotional patterns.



Inspiration: The TIM-NET framework and our Adaptation

Key TIM-Net Concepts We Adapted for our MFCC Branch:

- Temporal-Aware Blocks (TABs):
    1. Dilated Causal Convolutions (DC Conv): We implemented 1D dilated causal convolutions to expand the receptive field and model temporal dependencies without future data leakage.
    2. Core Block Structure: Our blocks also include Batch Normalization, ReLU activation, and Spatial Dropout, similar to TIM-Net's TABs.
- Bi-Directional Processing:
    1. We processed MFCC sequences in both forward and time-reversed directions to capture contextual information from past and "future" (reversed past) segments.
    2. Our outputs from these passes were then summed before global pooling.

Inspiration: The TIM-NET framework and our Adaptation

Simplifications & Differences in Our Implementation:

- Attention Mechanism: Our adaptation does not include the explicit sigmoid-based temporal attention mechanism (A(F_j-1)) within each Temporal-Aware Block as shown in the original TIM-Net.

- Dynamic Multi-Scale Fusion: The original TIM-Net fuses features (g_j) from multiple distinct temporal scales. Our MFCC branch processes the entire sequence through its stacked blocks and then performs a single global average pooling step on the combined bi-directional output, rather than fusing intermediate scale-specific features.

- Ensemble Component: Crucially, this adapted TIM-Net-inspired module serves as one branch of our larger ensemble model, working alongside a separate branch for Wav2Vec2 embeddings. The original TIM-Net is a standalone model.

Evaluation Metrics

To assess the performance of our Speech Emotion Recognition model, we employed the following standard evaluation metrics:

1. Accuracy:
   - The proportion of correctly classified emotion samples out of the total number of samples.
   - Calculated as: (True Positives + True Negatives) / (Total Samples)
   - While common, it can be misleading for imbalanced datasets.
2. Confusion Matrix:
   - A table visualizing the performance of the classification model.
   - Each row represents the instances in an actual class, while each column represents the instances in a predicted class.
   - Helps identify which emotions are often confused with each other.
3. Classification Report (from Scikit-learn):
   - Provides a detailed breakdown of key metrics per emotion class:
     - Precision: The ability of the classifier not to label a negative sample as positive (TP / (TP + FP)).
     - Recall (Sensitivity): The ability of the classifier to find all the positive samples (TP / (TP + FN)). Crucial for understanding how well each individual emotion is detected.
     - F1-Score: The weighted average of Precision and Recall (2 * (Precision * Recall) / (Precision + Recall)). Useful for a balanced measure, especially with class imbalance.
     - Support: The number of actual occurrences of the class in the dataset.

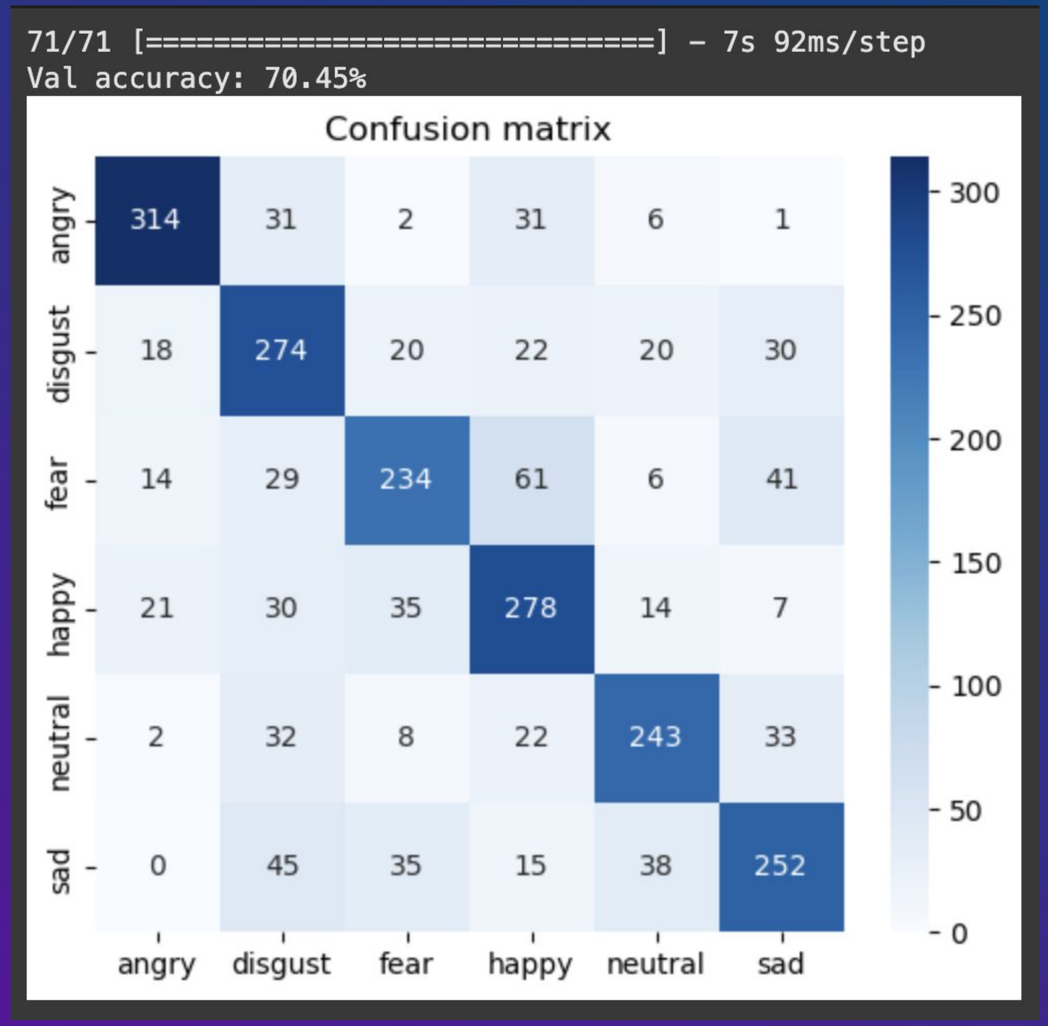
Experimental Results & Analysis

Our ensemble model demonstrated significant improvement over a baseline configuration:

1. Performance Metrics (on Combined Validation Set - 6 Emotions):

- Baseline (MFCC Branch - Simplified TIM-Net Inspired - Alone):
  - ❑ Accuracy: ~51%
- Proposed Ensemble Model (MFCC Branch + Wav2Vec2 Branch):
  - ❑ Overall Validation Accuracy / Weighted Average Recall (WAR): 70.45%

2. Confusion Matrix Analysis (Validation Set - 70.45% Accuracy):



Challenges Encountered

Throughout this project, we navigated several inherent challenges common in Speech Emotion Recognition:

1. Dataset Imbalance & Heterogeneity:
    - Varying Emotion Distributions: Combining RAVDESS, CREMA-D, TESS, and SAVEE resulted in an imbalanced distribution of the 6 target emotions (Neutral, Happy, Sad, Angry, Fearful, Disgust). Some emotions had significantly more samples than others.
    - Diverse Recording Conditions: Each dataset was recorded under different conditions (e.g., actors vs. crowd-sourced, different microphones, noise levels), introducing variability that the model had to generalize across.
2. Variability in Emotional Expression:
    - Inter-Speaker Differences: Speakers across the datasets (and even within) expressed emotions with unique vocal characteristics (accent, pitch, intonation), making it difficult to find universal emotional cues.
    - Acoustic Overlap: Some emotions (e.g., Sad and Neutral, or Fear and Sad as seen in confusion matrix) share similar acoustic properties, leading to misclassifications.
3. Subjectivity in Emotion Labeling:
    - The ground truth labels in these datasets, while curated, inherently involve a degree of human subjectivity. What one annotator perceives as "Sad," another might perceive as "Neutral" or a low-intensity "Fear." This can introduce noise into the training data.

Future work and discussion

Building on our current ensemble model, several exciting avenues for future work and enhancements can be explored:

- Refining the Temporal Modeling Branch (MFCCs):
  - Full TIM-Net Implementation: Explore implementing the complete TIM-Net architecture for the MFCC branch, including the original per-block sigmoid attention mechanism and the dynamic multi-scale fusion strategy, to potentially capture more nuanced temporal dynamics.
  - Alternative Architectures: Investigate other advanced temporal models like Transformers or different RNN variants (e.g., LSTMs with attention) for processing MFCC sequences.
- Enhancing Wav2Vec2 Integration:
  - Fine-tuning Wav2Vec2: Instead of using only static mean-pooled embeddings, fine-tune the pre-trained Wav2Vec2 model on the emotional speech data. This could adapt its representations more specifically to the SER task.
  - Attentive Pooling/Temporal Processing of Wav2Vec2 Outputs: Explore using attention mechanisms or lightweight temporal layers (e.g., 1D CNNs, RNNs) directly on the full sequence of Wav2Vec2 hidden states, rather than just mean-pooling, to retain more temporal information from these rich embeddings.
- Addressing Class Imbalance & Confusions:
  - Advanced Data Augmentation: Employ more sophisticated audio augmentation techniques specifically targeted at under-represented emotion classes.
  - Loss Function Modifications: Investigate loss functions designed for imbalanced classification (e.g., Focal Loss, class-weighted cross-entropy) to improve performance on minority classes and reduce specific confusions (like Sad/Neutral).
- Expanding Scope & Modalities:
  - Cross-Corpus Generalization & Domain Adaptation: Systematically evaluate and improve model generalization to unseen datasets or recording conditions using domain adaptation techniques.
  - Incorporating Additional Emotions: Re-integrate and develop strategies to effectively model "Calm" and "Surprise," which were filtered out in the current study.
  - Multimodal Emotion Recognition: Extend the framework to incorporate visual cues (facial expressions from RAVDESS/CREMA-D) or textual information (if available) for a more holistic emotion understanding.
  - Real-time SER: Investigate model optimization and architectural changes for deployment in real-time applications.

Reference

Jiaxin Ye, Xin-cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan, "Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1-5. DOI: 10.1109/ICASSP49357.2023.10096370.

Thank You

Any question…