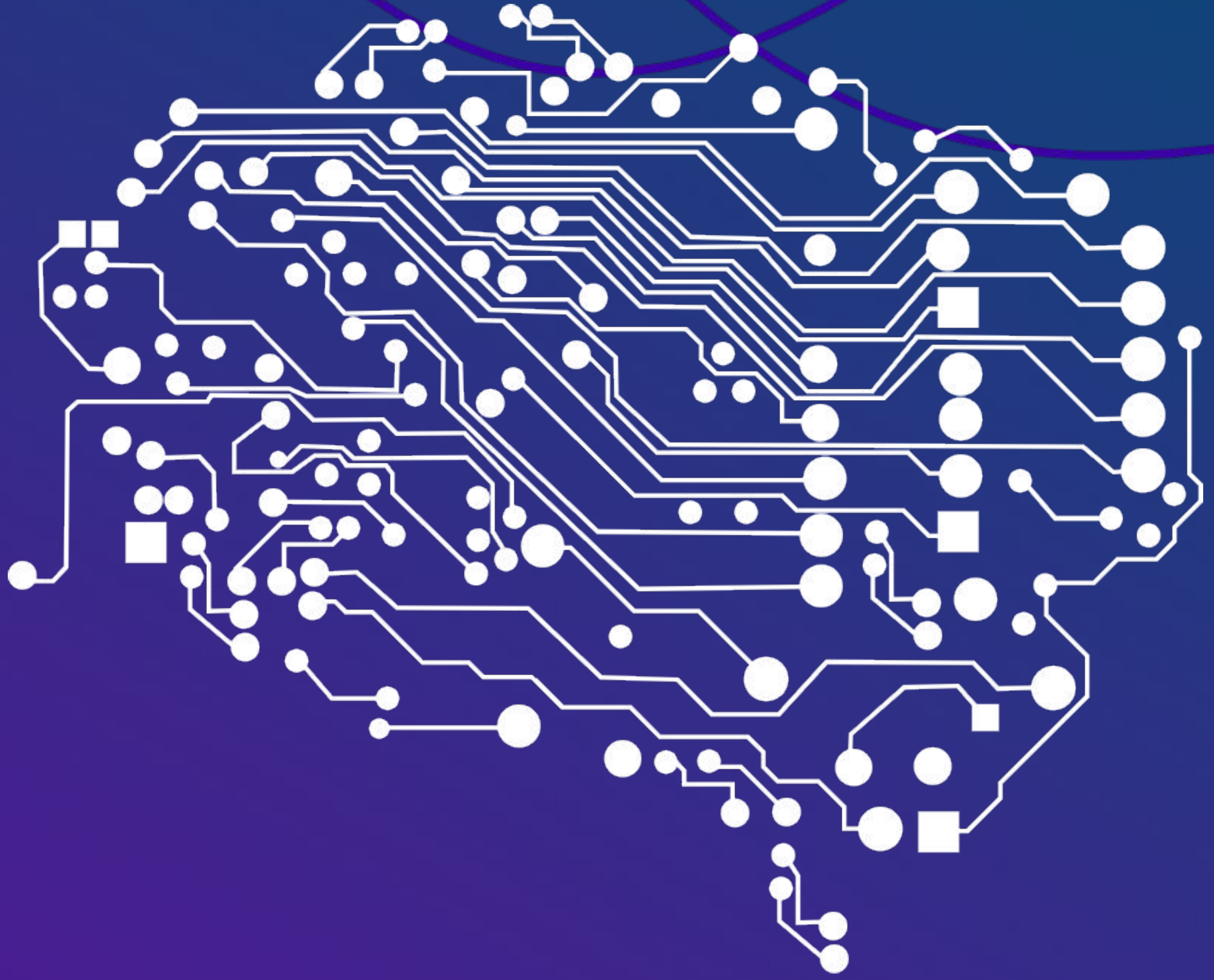
Speech Emotion Recognition (SER)

Vaibhav Kumar

Kanhaiya Kumar Sahu

Karan Deep Das

Chinmay Thakur



INTRODUCTION

Speech emotion recognition detects emotions from speech signals

Used in customer support, healthcare, and Public assistance

Problem formulation

AI has the potential to revolutionize numerous industries by providing new and innovative solutions to complex problems. From healthcare to finance, AI is already being used to improve efficiency, increase accuracy, and drive innovation. This section will explore some of the practical applications of AI in various industries, demonstrating the real-world impact of this technology.

Challenge: High variability across speaker ,context and noise
Goal: Maximize classification accuracy under resource constraints

EaTemp

Database used

01

RAVDESS

27 actors,8 emotions

02

TESS

7 emotions, Female speaker

03

CREMA-D

91 actors, various emotions.

04

SAVEE

4 male actors,7 emotion

Tools and libraries

Programming language

python

Libraries

Librosa , scikit-learn, seaborn , pandas , matplotlib

Deep learning framework

Keras

Feature Extraction

- Used MFCC as main audio Features
- Mean-pooled wav2vec2 embedding trunk
- Visual features to analyze pattern

Proposed methodology

Combined all datasets into a unified structure

Applied Efficient mean-pooling to reduce memory footprint.

integration of TIM-Net temporal blocks with wav2vec2 embeddings

Split Achieves >70 % accuracy on combined emotional speech dataset

Evaluation Metrices

Used Accuracy , confusion Matrix , Classification Report

Visualized a confusion matrix to analyse performance

Results

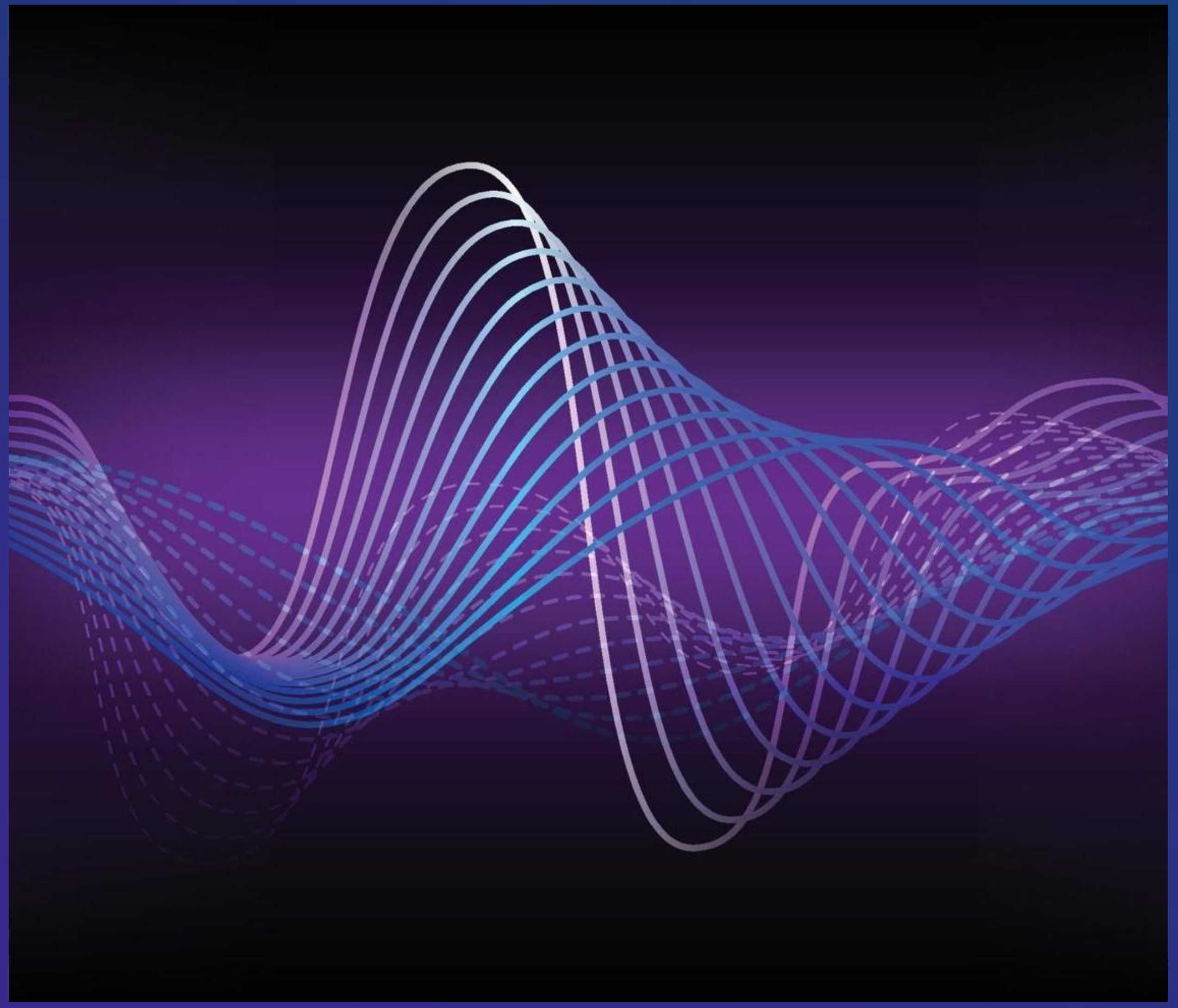
Baseline TIM alone ~ 51% accuracy

Ensemble model : ~70-71 % validation accuracy

Confusion matrix analysis highlights improved class separability but confusion in some emotions like fear , Sad neutral

Challenges Faced

An imbalanced dataset and different sets of emotions across emotions and datasets



The EU's Variability in speaker accents and noise. Subjectivity in labelling emotional tones

Future work and discussion

Use RNN or CNN for better temporal modelling. Explore multilingual and real-time SER. Incorporate visual signals (multimodal emotion recognition)

Reference

TEMPORALMODELINGMATTERS: A NOVELTEMPORALEMOTIONALMODELING APPROACH FOR SPEECHEMOTIONRECOGNITION. Jiaxin Ye, Xin-cheng Wen, Yujie Wei, Yong Xu3, Kunhong Liu, Hongming Shan1. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096370

Thank You

Any question…