Estimating Emotions from Speech signals for Emotion-Preserving Translation

Bagiya Lakshmi S (21011101029)
Chunduri Suhasini (21011101036)
Cynddia Balamurugan (21011101037)
Harini C J (21011101045)

Dr. G. Anushiya Rachel
Shiv Nadar University Chennai

SHIV NADAR
UNIVERSITY
CHENNAI

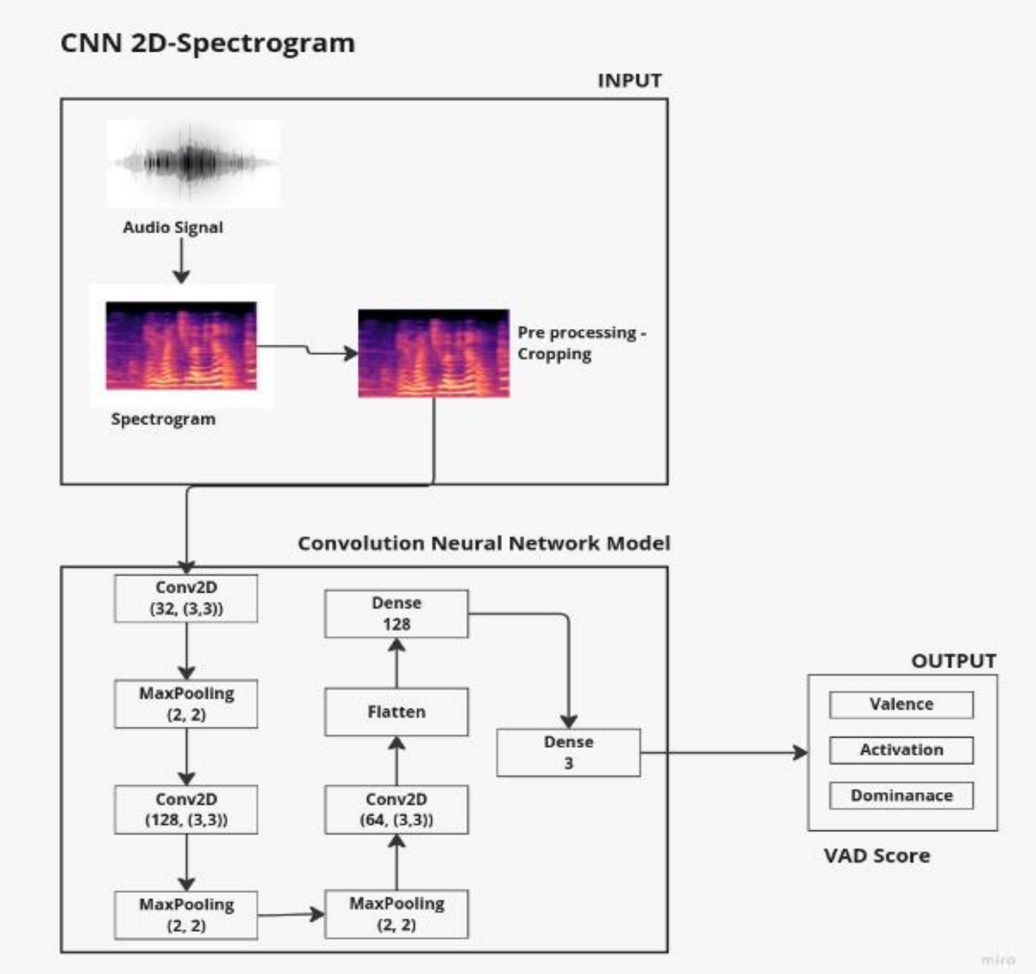
Objectives:

- To build an emotion detection model using acoustic features like MFCC, Chroma, Spectral Contrast, Zero Crossing Rate, and spectrograms.
- Spectrograms will be used to build CNN-based models for more accurate emotion detection.
- To predict the emotional dimensions of Valence, Arousal, and Dominance (VAD) from speech signals.
- To demonstrate how these VAD scores can be used to influence emotion-aware translation through integration with a pretrained model.

Overview:

This project introduces the emotion detection system that integrates acoustic feature extraction from audio signals to identify emotions. The primary objective of the system is to detect emotions, specifically by predicting Valence, Arousal, and Dominance (VAD) scores. These emotional dimensions are then utilized for various applications, such as emotion-aware systems or multilingual tasks. While emotion preservation is the core goal, a pretrained machine translation model is employed to demonstrate how emotional states (Valence, Arousal, Dominance) from the emotion detection model can influence the translated speech.

Model Development:



CNN 2D-Spectrogram

INPUT

Audio Signal

Spectrogram

Pre processing - Cropping

Convolution Neural Network Model

Conv2D (32, (3,3))

MaxPooling (2, 2)

Conv2D (128, (3,3))

MaxPooling (2, 2)

MaxPooling (2, 2)

Conv2D (64, (3,3))

Flatten

Dense 128

Dense 3

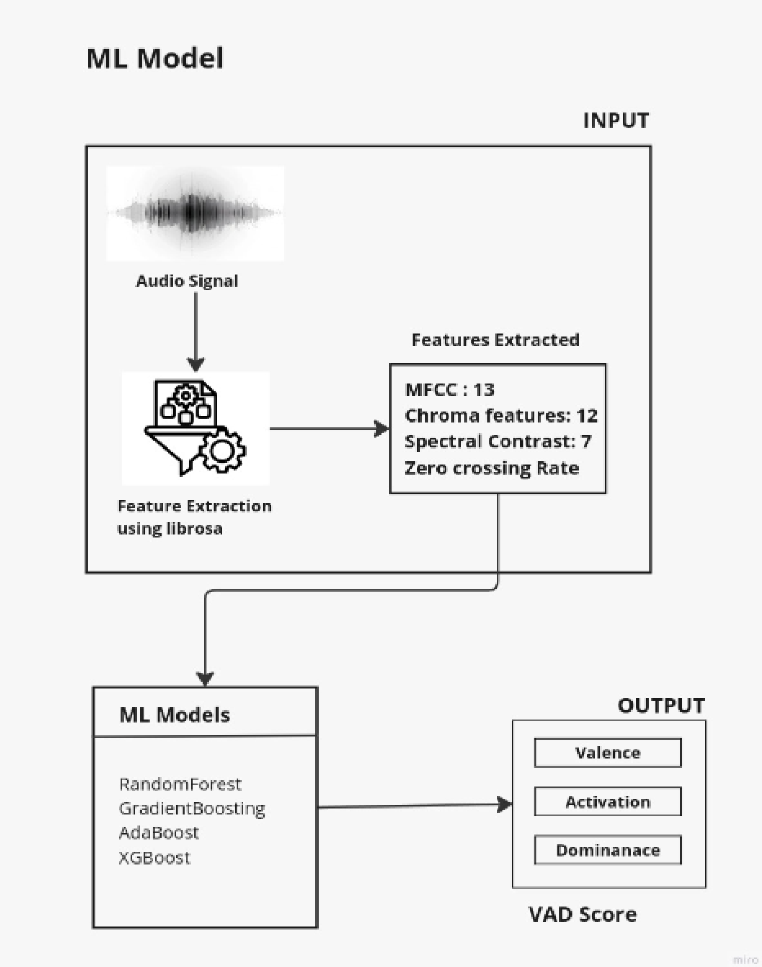
OUTPUT

Valence

Activation

Dominanace

VAD Score



ML Model

INPUT

Audio Signal

Feature Extraction using librosa

Features Extracted

MFCC : 13
Chroma features: 12
Spectral Contrast: 7
Zero crossing Rate

ML Models

RandomForest
GradientBoosting
AdaBoost
XGBoost

OUTPUT

Valence

Activation

Dominanace

VAD Score

SHIV NADAR
UNIVERSITY
CHENNAI

Workflow: Emotion Detection to Translation

Step 1: **Feature Extraction**

Extract acoustic features (MFCC, Chroma, Spectral Contrast, Zero Crossing Rate) from the audio input.

Step 2: **Emotion Detection**

For 2D CNN (Spectrogram-based model) we generated a 2D spectrogram from the audio signal.

A 2D CNN was used to detect emotions based on the spectrogram features. For Other Models (e.g., RNN, Random Forest) extracted features such as (MFCC, Chroma, Spectral Contrast, Zero Crossing Rate) were used to predict Valence, Arousal, and Dominance (VAD) using machine learning or deep learning models.

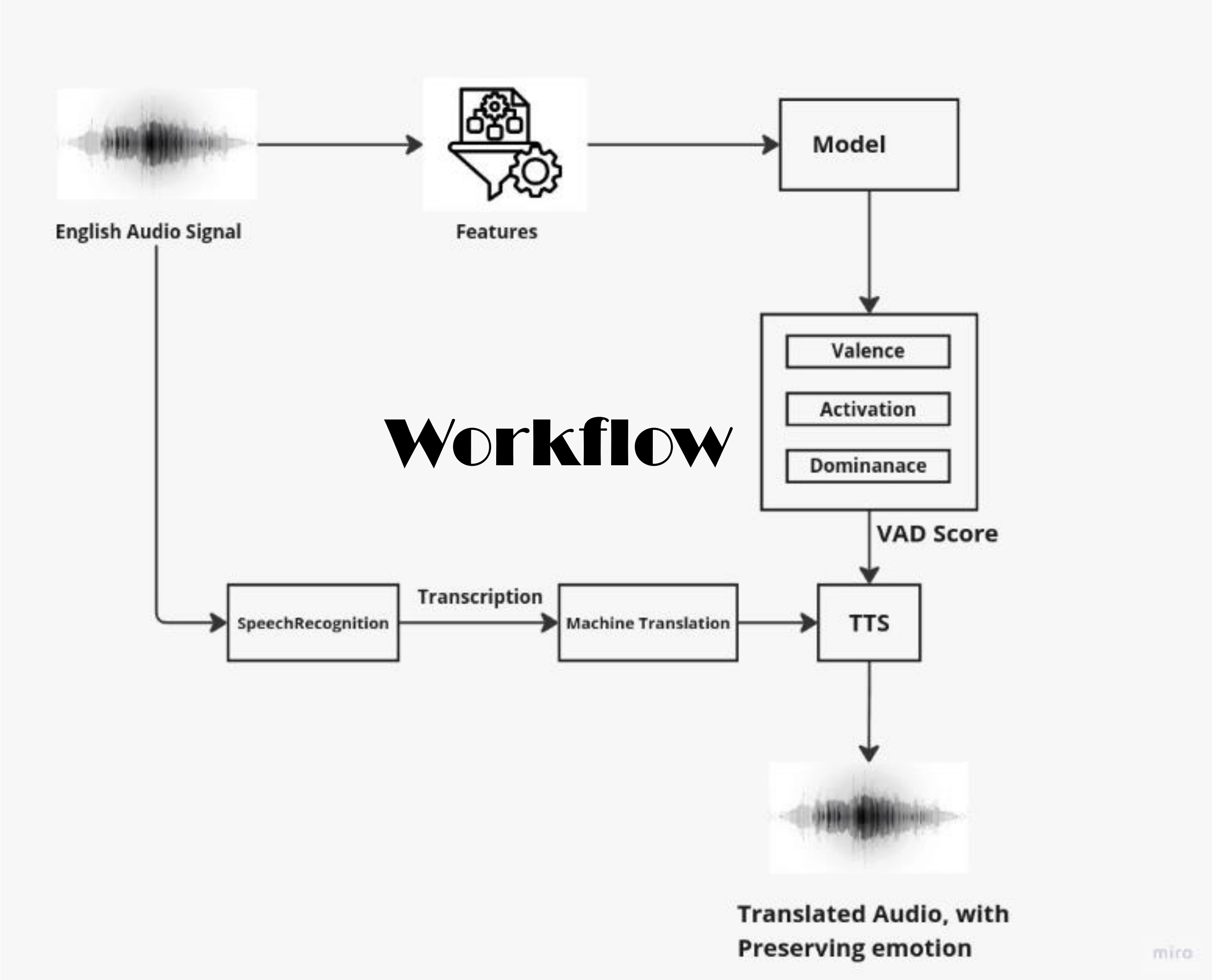
Step 3: Incorporate VAD Scores into Translation Model

Integrated the predicted VAD scores (Valence, Arousal, and Dominance) as input into a T5-based translation model to adjust the emotional tone for the translation.

Step 4: Generate Emotion-Preserved Translation

Used GTTS (Google Text-to-Speech) to generate the translated speech in the target language, ensuring that the detected emotional tone is preserved in the output.

Workflow

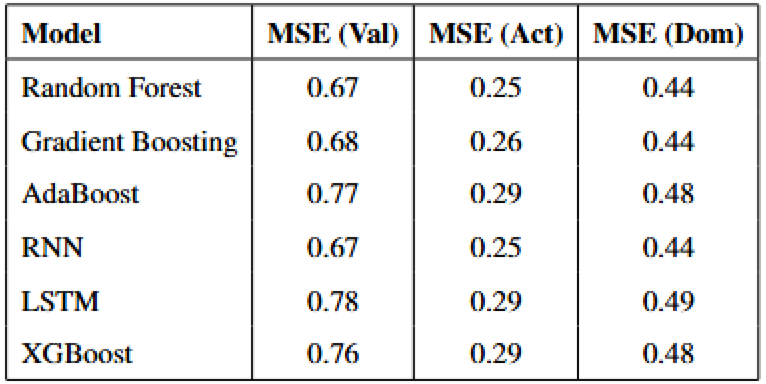


SHIV NADAR
UNIVERSITY
CHENNAI

RESULT ANALYSIS:

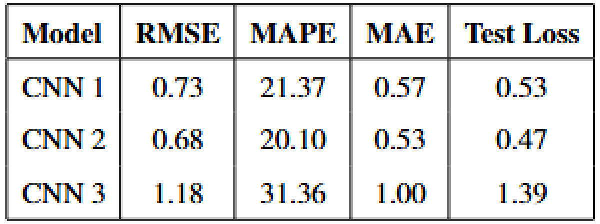
- **RNN and Random Forest** achieved the lowest **MSE values** across emotional dimensions, making them the most effective for emotion detection.
- **CNN 2** showed the best results among spectrogram-based models, with the lowest **RMSE** and **MAPE**.
- The **emotion-preserving translation system** successfully maintained the **emotional tone** during translation by leveraging **VAD scores** predicted by the top-performing models.

Model Comparison



Model	MSE (Val)	MSE (Act)	MSE (Dom)
Random Forest	0.67	0.25	0.44
Gradient Boosting	0.68	0.26	0.44
AdaBoost	0.77	0.29	0.48
RNN	0.67	0.25	0.44
LSTM	0.78	0.29	0.49
XGBoost	0.76	0.29	0.48

Table 5.1: MSE for Machine Learning and Deep Learning Models.



Model	RMSE	MAPE	MAE	Test Loss
CNN 1	0.73	21.37	0.57	0.53
CNN 2	0.68	20.10	0.53	0.47
CNN 3	1.18	31.36	1.00	1.39

Table 5.2: Evaluation Metrics for CNN Models.

CONCLUSION:

Summary of Findings

The project successfully demonstrates how emotion-preserving speech translation can be achieved by combining emotion detection with multilingual translation systems. The models based on MFCC, Chroma, Spectral Contrast, and Zero Crossing Rate performed well in predicting VAD scores, which were integrated into the translation pipeline.

Future Scope

Future work could focus on improving the emotional accuracy of translations by experimenting with different feature extraction techniques. The plan is to keep the model constant while exploring a variety of features for better emotion detection. Additionally, the system could be expanded to support more languages and applied to real-time speech analysis for speech-to-speech emotion-preserving translation tasks. This would further enhance the system’s adaptability and performance in diverse real-world scenarios.