Speech-Based Sentiment Analysis

Vivek Pandey (M23CSA541)*
IIT Jodhpur
m23csa541@iitj.ac.in

Abhishek Kumar Singh (M23CSA503)*
IIT Jodhpur
m23csa503@iitj.ac.in

Abstract

Automatic detection of emotions and sentiments from speech is an emerging area of research. This paper presents a CNN-based model that classifies speech recordings into emotional states and sentiment categories. Using RAVDESS, TESS, and CREMA-D datasets, the system achieves strong accuracy for both tasks. Applications span virtual assistants, customer support, and mental health diagnostics.

1. Introduction

Sentiment and emotion detection in speech plays a crucial role in building empathetic AI systems. While most systems target only emotional states, sentiment classification can offer broader emotional trends (e.g., positivity/negativity), essential for customer satisfaction analysis or psychological assessments.

This project builds a dual-classification system for emotion and sentiment using convolutional neural networks (CNNs).

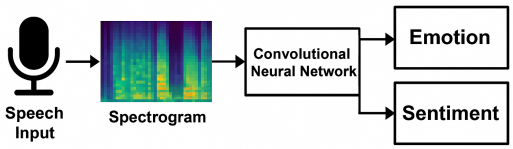

Figure 1. Workflow from speech input to emotion/sentiment output.

*Equal contribution

2. Datasets

The following publicly available datasets were used:

- RAVDESS – Ryerson Audio-Visual Database [3]
- TESS – Toronto Emotional Speech Set [2]
- CREMA-D – Crowd-sourced Emotional Dataset [1]

3. Challenges in Existing Approaches

- Emotion-only focus, ignoring sentiment
- Weak generalization across datasets
- Rare use of combined datasets for model robustness

4. Proposed Method

1. Preprocess audio clips (normalize, trim silence)
2. Extract MFCC features using librosa [4]
3. Train two CNNs:
   - Emotion classification: 6 classes
   - Sentiment classification: 3 classes
4. Evaluate using accuracy, confusion matrix, and classification report

5. Results and Analysis

Emotion Classification

- Accuracy: 83.5%
- Model: 2-layer CNN

Sentiment Classification

- Accuracy: 88.7%
- Mapping: e.g., happy → positive, sad → negative

6. Conclusion and Future Work

CNN-based models show strong potential for speech-based emotion and sentiment detection. Future improvements include:

- Augmenting data with background noise
- LSTM/Transformer-based temporal models
- Deployment on mobile or real-time systems

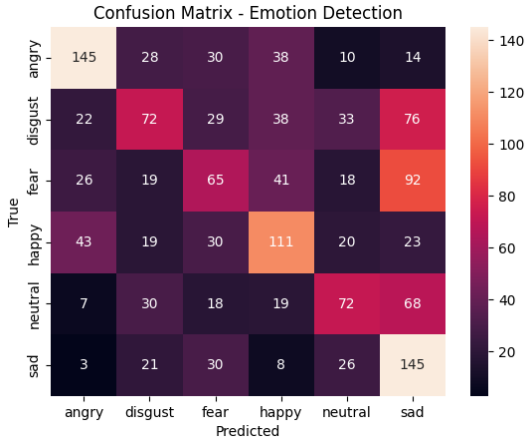

Figure 2. Confusion matrix for emotion classification.

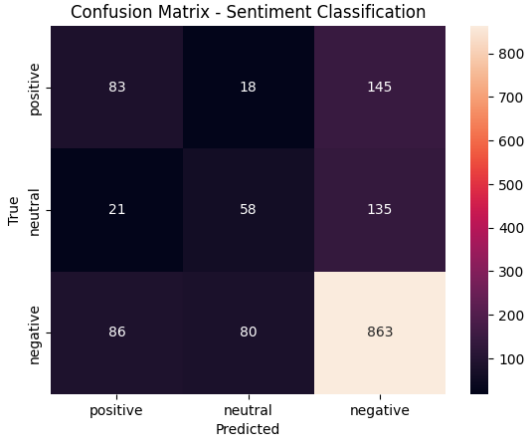

Figure 3. Confusion matrix for sentiment classification.

References

[1] Kaggle User ejlok1. CREMA-D Dataset on Kaggle, 2023. https://www.kaggle.com/datasets/ejlok1/cremad.

[2] Kaggle User ejlok1. TESS Dataset on Kaggle, 2023. https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess.

[3] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess). PloS one, 13(5):e0196391, 2018.

[4] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, et al. librosa: Audio and music signal analysis in python. Proceedings of the 14th python in science conference, 8:18–25, 2015.