Speech Understanding

Niket Agrawal & Ritesh Lamba

Computer Science and Engineering Department

IIT Jodhpur

February 2025

- Two Approach Available
    - Traditional Machine Learning Models
    - Deep Learning Models

Features

- MCFF
- Chroma
- Special Contrast

Strengths

- Simple and Interpretable
- Effective for small dataset

Limitations

- Limited ability to capture complex patterns in audio data
- Requires manual feature engineering

CNNs (Convolutional Neural Networks)

- Use Mel-Spectrograms or spectrograms as input

Strengths

- Captures spatial patterns in spectrograms effectively

Limitations

- Struggles with temporal dependencies in speech

RNNs/LSTMs (Recurrent Neural Networks/Long Short-Term Memory)

- Process sequential data (e.g., MFCCs over time)

Strengths

- Handles temporal dependencies well

Limitations

- Computationally expensive; prone to vanishing gradients

Transformer-Based Models

- Examples: Wav2Vec 2.0 fine-tuned for emotion recognition

Strengths

- Captures long-range dependencies; highly accurate

Limitations

- Requires large datasets and computational resources