AI CLUB Task: Speech Emotion Recognition (SER) using CNNs

Submission Deadline: 5 Feb 2025

Project Overview: This task focuses on the intersection of digital signal processing and computer vision. By treating audio frequency maps as images, you will build a system capable of identifying human sentiment from 3-second speech samples.

Dataset & Objective

You will use the **RAVDESS** dataset, which contain professional recordings of actors expressing specific emotions. Your objective is to build a 2D Convolutional Neural Network (CNN) that can generalize across different voices, pitches, and genders.

Target Classes: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised.

Phase 1: Preprocessing & EDA

- **Audio Cleaning:** Implement silence trimming to remove dead air.
- **Visual Analysis:** Compare Mel-spectrograms of high-arousal emotions (Angry) vs. low-arousal (Sad) to identify "visual" differences in spectral energy.
- **Feature Engineering:** Convert raw waveforms into Log-Mel Spectrograms, ensuring all outputs are padded to a uniform shape
- **Data Augmentation:** Because the dataset is small, you must apply techniques like **Noise Injection**, **Pitch Shifting**, or **Time Stretching** to the training set to prevent the model from simply "memorizing" the actors' voices.

Phase 2: Architecture & Training

- **The Model:** Construct a 2D CNN using multiple convolutional blocks (Conv2D, BatchNormalization, and Pooling).
- **Regularization:** Utilize Dropout and experiment with **Global Average Pooling** to minimize overfitting.
- **Validation:** Perform a stratified split (80% Train, 10% Val, 10% Test) to ensure equal representation of all 8 emotions.
- **Evaluation:** Beyond standard accuracy, report the **Macro F1-Score** and generate a **Confusion Matrix**. Analyze if the model shows a "Pitch Bias" by testing performance differences between male and female speakers.

Phase 3: Deliverables

- **Notebook:** An .ipynb file containing your EDA, training curves, and evaluation metrics.
- **Model Weights:** The saved weights file (.h5, .pth, or .keras) of your best-performing iteration.
- **Live Inference:** A predict.py script that accepts an unseen .wav file, processes it, and prints the predicted emotion with a confidence percentage.

Reference Links & Resources

Datasets:

- **RAVDESS:** Kaggle - Audio-Visual Database of Emotional Speech and Song

Core Topics:

- **Audio Processing:** Librosa Documentation & Tutorials
- **Technical Guide:** Speech Emotion Recognition using CNN (Researchgate)
- **CNN:** Neural Networks Lecture 5 CNN

Mel Spectrograms Explained Easily

This video provides a clear, intuitive breakdown of how audio signals are transformed into Mel-spectrograms, which is the foundational preprocessing step for your CNN model.