PROJECT ABSTRACT

Speech Emotion Recognition using Machine Learning

Speech carries more than words: tone, pitch and rhythm reveal the speaker's emotional state.

01 Title

Speech Emotion Recognition using Machine Learning

02 Introduction / Background

Speech carries more than words: tone, pitch and rhythm reveal the speaker's emotional state. Machines reading these cues power call-center analytics, driver monitoring and mental-health screening. Yet most student AI projects stop at image classification. Speech Emotion Recognition (SER) is a rarer, more defensible final-year build.

03 Problem Statement

Emotion in speech hides in subtle acoustic variations that simple thresholds cannot capture. This project builds an end-to-end system converting voice recordings — uploaded or recorded live — into accurate emotion labels with honest confidence scores: standardized preprocessing, industry-standard acoustic features, three classifiers compared on the recognized RAVDESS benchmark, served through a web app anyone can test.

04 Objectives

- Preprocess audio to 22.05 kHz: trim silence, normalize amplitude.
- Extract 40 MFCCs + deltas with librosa (40 × 216 per 3-second clip).
- Benchmark a 2D CNN, an LSTM and an SVM on RAVDESS (1,440 clips, 24 actors, 8 emotions).
- Report accuracy, confusion matrices, per-class F1 per model.
- Deliver VoiceMood: Streamlit app with upload, mic recording, confidence radar, emotion timeline.

05 Proposed System / Working

Audio is resampled to 22.05 kHz, silence-trimmed and normalized with librosa. Forty MFCCs plus deltas — mimicking human hearing — form fixed 216-frame matrices. The dataset is RAVDESS: 1,440 acted clips across eight emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised). A 2D CNN (Conv2D + BatchNorm + Dropout) learns spectro-temporal patterns; an LSTM captures prosodic dynamics; an SVM baseline uses aggregated MFCC statistics. Built in TensorFlow/Keras, served via VoiceMood (Streamlit), re-trainable with a Colab GPU notebook.

The CNN reaches roughly 70–75% on the RAVDESS held-out split — an honest number examiners respect. Confusion matrices and per-class F1 show where models confuse similar emotions (calm vs neutral) and quantify the deep models' edge over the SVM baseline. In the live VoiceMood demo, an examiner records their own voice and instantly sees the predicted emotion, confidence radar and emotion timeline.

06 Technologies / Components Used

Python 3.10 · NumPy · pandas · librosa · SoundFile · TensorFlow/Keras · scikit-learn · Streamlit · Matplotlib/Plotly · RAVDESS dataset trained model included · dataset source linked in docs

07 Key Features

- Upload or record voice clips (WAV/MP3) in the browser.
- librosa MFCC + delta pipeline — the standard in SER research.
- CNN trained on RAVDESS; LSTM and SVM variants for comparison.
- Confidence radar per prediction; emotion timeline for longer clips.
- Batch prediction mode; CPU real-time inference.

08 Applications / Use Cases

- Call-center experience monitoring
- in-car drowsiness and road-rage detection
- mental-health screening assistants
- assistive communication tools

10 Conclusion

This project delivers a complete, benchmarked SER system — from waveform to web app — with the comparative rigor viva examiners probe for. Signal-processing fundamentals, three evaluated model families and honest limitations (acted-data bias, short-clip reliability) make it a defensible audio-AI build. VoiceMood makes it tangible: record, predict, see emotion in sound.