Speech Emotion Recognition using Deep Learning

Abstract

The fast-expanding discipline of speech emotion recognition (SER) uses machine learning and signal processing to identify human emotions in speech. An end-to-end SER system is shown in this research, which combines several open-source datasets, extracts prosodic and spectral information, and trains hybrid deep learning models for reliable classification. A Streamlit-based user interface is used to further deploy the system, enabling real-time prediction utilizing live microphone recordings or uploaded audio files. The hybrid model performs well, according to the results, making it a promising tool for education, healthcare, and human–computer interaction.

Problem Statement

In addition to words, human communication also includes prosody, pitch, and tone, all of which express emotions. Automating voice recognition of emotions can improve assistive technologies, call center analytics, mental health monitoring, and human–computer interaction. However, the necessity for robust feature extraction, uneven labeling, background noise, and dataset fluctuations provide difficulties. By merging several datasets, using feature engineering methods, and creating hybrid deep learning models that capture both spectral and prosodic information, this effort tackles these problems.

Tools & Technologies Used

- Programming Language: Python
- Libraries & Frameworks: Pandas, NumPy, Librosa, Parselmouth, Scikit-learn, TensorFlow/Keras, Matplotlib, Streamlit
- Datasets: RAVDESS, CREMA-D, TESS, EmoDB, IESC
- Artifacts Management: Joblib, NumPy

Methodology

1. Dataset Preparation: Multiple datasets (RAVDESS, CREMA-D, TESS, EmoDB, IESC) were merged, labels standardized, and final cleaned dataset saved.
2. Feature Extraction: Spectrogram features (Mel, MFCC, Chroma) and prosodic features (Pitch, Energy, Formants, Jitter, Shimmer) were extracted with augmentations (Pitch shift, Time-stretching, Noise injection, SpecAugment).
3. Preprocessing: Per-channel normalization of spectrograms, standardization of prosodic features, and label encoding.

4. Model Training: CNN model trained on spectrograms and hybrid CNN+MLP model trained on both spectrogram and prosodic features.
5. Inference & Deployment: Streamlit app deployed with options to upload WAV files or record audio via microphone, providing predictions with probability visualizations.

Results

- Training Accuracy: 79%
- Testing Accuracy: 73.89%
- Macro-averaged F1-score: 76%

Combining spectral and prosodic data improves accuracy, as seen by the hybrid model's superior performance over the CNN-only method. For interpretability, the Streamlit interface effectively offers probability visualizations and real-time predictions.

Conclusion & Future Scope

This experiment shows how well spectral and prosodic information may be used to provide strong speech emotion recognition. The initiative closes the gap between research and real-world applications by establishing a real-time interface and combining multiple datasets.

Future work can include:
- Extending to more diverse datasets with multilingual support.
- Exploring advanced models like Transformers and wav2vec for end-to-end learning.
- Integrating multimodal inputs (speech + facial expressions) for richer emotion recognition.
- Optimizing models for mobile and edge deployment in real-world applications.