Speech Emotion Recognition using Convolutional-Recurrent Hybrid Neural Networks

Mahesh Divakaran Namboodiri, Samudyata Sudarshan Jagirdar, Sayantika Paul

Department of Electrical, Computer and Energy Engineering, Arizona State University, 1151 S Forest Ave Tempe, AZ 85281

Ira A. Fulton Schools of Engineering
Electrical, Computer and Energy Engineering

Abstract

SER MODEL

Speech Emotion Recognition (SER) enhances human-computer interaction by enabling machines to interpret emotions from speech, offering richer and more intuitive communication. This study reviews SER methodologies, focusing on modern deep learning approaches like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN-RNN models. These architectures combine spatial feature extraction and temporal dependency modeling to tackle emotion recognition complexities.

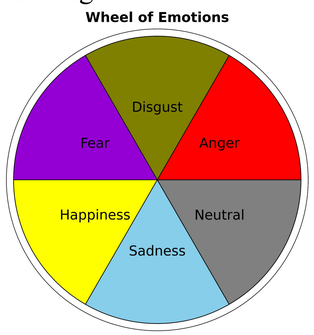
Key datasets, including RAVDESS, TESS, and CREMA-D, are analyzed for their role in improving model generalization across diverse emotional states. Challenges such as speaker variability, noise resilience, and real-time processing are highlighted, alongside applications in mental health diagnostics and customer service. This work provides a foundation for advancing SER systems through future research.

Introduction

Speech Emotion Recognition (SER) is vital in human-computer interaction and affective computing. However, accurately detecting emotions from speech remains challenging due to the complexity of emotional expression and individual variability. This research tackles these challenges by:

- Implementation of Robust Feature Extraction Techniques
- Development of CNN+RNN Hybrid Model
- Application of Data Augmentation Strategies

- Six emotions considered.
- Dataset: RAVDESS, TESS, and CREMA-D.
- 10,898 samples of audio data used.



Related Work

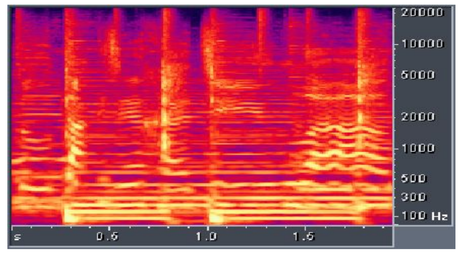
- CNNs: Excel at sound classification using spectrograms, outperforming traditional DNN models.
- RNNs: Specialize in capturing temporal patterns and identifying emotional features in sequential speech data.
- CNNs + RNNs: Chen et al. demonstrate improved emotion recognition, especially for sadness, using a 3D CRNN with attention, combining 3D convolutions and LSTM units.

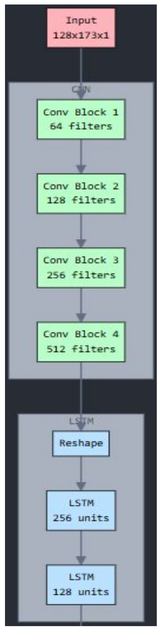
References

1. K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Convolutional recurrent neural networks for music classification,”
2. K. J. Piczak, “Environmental sound classification with convolutional neural networks”

Data Pre-processing and Architecture

- Advanced Signal Processing : 22,050 Hz sampling rate
- Mel-Spectrogram Generation : Employed 128 mel frequency bands for detailed frequency resolution
- Feature Normalization
- Time stretching
- Padding/truncation





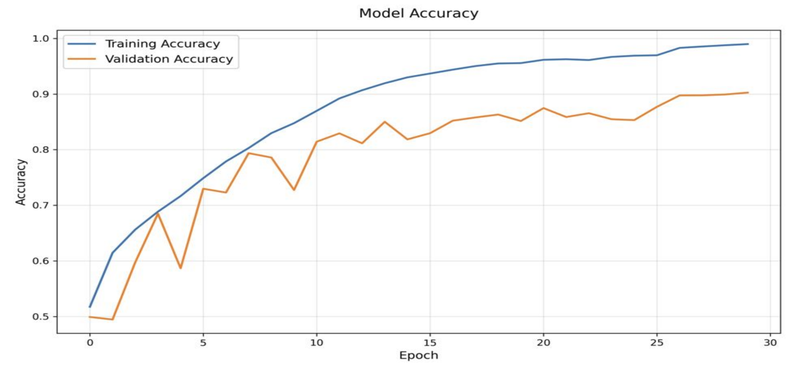
- Audio signals as 2D frequency-time data, analyzed by CNNs..
- Created by stacking FFTs of audio segments, with intensity shown as brightness.
- Models human speech in a graphical representation for analysis.

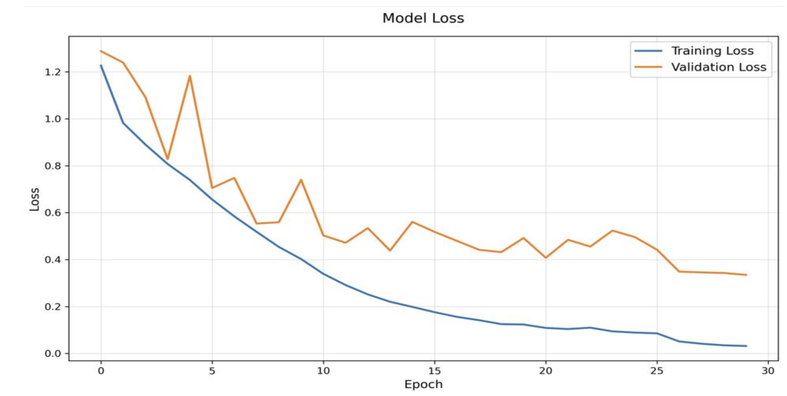
Training Characteristics

- Batch normalization
- Dropout layers
- Early stopping
- Learning rate reduction on plateau

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	0.001
Weight Decay	0.0001
Loss Function	Categorical Crossentropy
Metrics	Accuracy
Batch Size	32
Epochs	30
Early Stopping	Monitor 'val_loss', Patience 15, Restore Best Weights
Learning Rate Reduction	Monitor 'val_loss', Factor 0.1, Patience 5, Min LR 1e-6

Results





- **Training Accuracy:** 98.96%
- **Validation Accuracy:** 90.26%
- **Test Accuracy:** 90.26%
- Successful convergence after 30 epochs

Future Work:

- Addressing Dataset Limitations
- Emotion Complexity (Tackling subtle or mixed emotion)
- Use Mel-Frequency Cepstral Coefficients of the data to detect more subtle changes.