ABSTRACT

By

Reza Aurelio Brilliansah
21/475039/PA/20515

Speech Emotion Recognition (SER) is a challenging area of audio signal processing that seeks to automatically identify emotional states from human speech. This undergraduate thesis presents a SER system designed to classify eight discrete emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprise in English speech. The proposed approach integrates Mel spectrogram and MFCC coefficient feature extraction techniques to capture perceptually relevant audio characteristics. The system leverages a hybrid deep learning architecture that fuses Convolutional Neural Networks (CNNs) and Bi-Directional Gated Recurrent Units (Bi-GRUs) with attention mechanisms, enabling the model to effectively learn both spatial and temporal patterns in speech. To optimize model performance, Full Model Selection is performed with Particle Swarm Optimization, automating the search for optimal architectural configurations and hyperparameters. Experimental results, validated through stratified 5-fold cross-validation and ANOVA statistical analysis, demonstrate that the proposed fusion model achieves 86.05% ± 1.97% accuracy, outperforming several baseline and state-of-the-art SER systems while achieving comparable performance to real-time transformer-based approaches. Data analysis using CMDS and LDA provided insights into feature separability and enabled interpretation of per-class emotion performance. The final model processes audio samples with an average inference time of less than 30 ms, making it suitable for low-latency applications.

Keywords: Speech Emotion Recognition, Signal Processing, Deep Learning