Speech Emotion Recognition Project Report

1 Introduction

Speech Emotion Recognition (SER) is the task of identifying human emotions from spoken language. This project explores SER by extracting features from audio recordings and classifying them using a neural network.

2 Data Collection and Preprocessing

2.1 Dataset Description

The dataset comprises two primary directories: `Audio_Speech_Actors_01-24` and `Audio_Song_Actors_01-24`. Each directory contains subfolders for various actors, with each subfolder holding `.wav` files representing different emotional states.

2.2 Feature Extraction

We extracted three main types of features from the audio files:

- **MFCC (Mel-Frequency Cepstral Coefficients)**: These coefficients describe the power spectrum of the audio signal and are crucial for capturing the timbral aspects of speech.
- **Chroma Features**: These features reflect the 12 pitch classes and provide insight into the harmonic content of the audio.
- **Mel Spectrogram**: This feature represents the short-time power spectrum of frequencies and is useful for understanding the energy distribution over time.

2.3 Data Scaling and Encoding

To prepare the data for modeling, we standardized the features using a standard scaling technique, which normalizes the feature values to have a mean of zero and a standard deviation of one. Labels were encoded into a one-hot format to facilitate multi-class classification.

2.4 Data Splitting

The dataset was divided into training and test sets, with 80% of the data used for training the model and 20% reserved for evaluation. This split ensures that the model's performance can be assessed on unseen data.

3 Model Development

3.1 Model Architecture

We designed a neural network with a single hidden layer. The hidden layer uses the ReLU activation function to introduce non-linearity, and the output layer employs a softmax activation function to classify the emotions into distinct categories.

3.2 Model Compilation and Training

The model was compiled using the Adam optimizer and categorical cross-entropy loss function, which are standard choices for classification problems. It was trained for 200 epochs with a batch size of 256, allowing the model to learn from the data and adjust its parameters.

4 Results and Discussion

4.1 Model Performance

The model achieved a test accuracy of approximately 76.0%. This indicates that the model correctly classified emotions 76.0% of the time on the test set. The classification report highlights the precision, recall, and F1-score for each emotion class.

4.2 Analysis

- The model showed solid performance across most emotion classes, with particular strength in precision and recall for some categories.
- Some classes exhibited lower recall, suggesting areas where the model may struggle with distinguishing certain emotions.
- Overall, the accuracy indicates a robust model, though there is room for improvement by exploring advanced neural network architectures or additional features.

5 Conclusion

This project successfully developed a Speech Emotion Recognition system that achieved a test accuracy of 76.0%. The results validate the feature extraction and classification approach used. Future work should focus on enhancing the model's performance by experimenting with different neural network architectures and incorporating additional audio features.

6 Classification Report

```
              precision    recall  f1-score   support

           0       0.82      0.74      0.78        38
           1       0.78      0.80      0.79        81
           2       0.84      0.66      0.74        73
           3       0.67      0.83      0.74        71
           4       0.83      0.83      0.83        69
           5       0.83      0.74      0.78        80
           6       0.62      0.64      0.63        45
           7       0.64      0.79      0.71        34

    accuracy                           0.76       491
   macro avg       0.75      0.75      0.75       491
weighted avg       0.77      0.76      0.76       491
```