Speech Emotion Recognition(SER)

Speech Understanding Programming Assignment-2

Project Report - Q-2

MFCC Feature Extraction and

Comparative Analysis of Indian Languages

Prepared By-

- Shyam Vyas (M23CSA545)

1. Introduction

Automatic speech language identification is a critical application in the area of speech processing. In this project, we study the feature extraction of Mel-Frequency Cepstral Coefficients (MFCC) from speech data and use them to classify audio samples into various Indian languages. The dataset consists of multiple Indian-language audio recordings, and the objective is to find meaningful features from these audio files and use “Support Vector Machine (SVM)” for classification.

2. Objective

The primary objective of this project is to:

1. Extract **MFCC features** from audio files in different Indian languages.
2. Visualize the MFCC spectrograms for a comparative analysis.
3. Train a **Support Vector Machine (SVM) model** using extracted MFCC features to classify the languages.
4. Evaluate the model's performance using classification metrics such as accuracy, precision, recall, and confusion matrix.

3. Dataset

The dataset used for this project has audio files in **MP3 format**, separate folders for every language. The dataset was taken from Kaggle, **"Audio Dataset with 10 Indian Languages"**. For the purpose of this study, we selected the following three languages:

- **Hindi**
- **Tamil**
- **Bengali**

Each language folder contains multiple audio samples spoken by different speakers. The dataset provides sufficient diversity to examine how MFCC features differ across languages.

4. Methodology

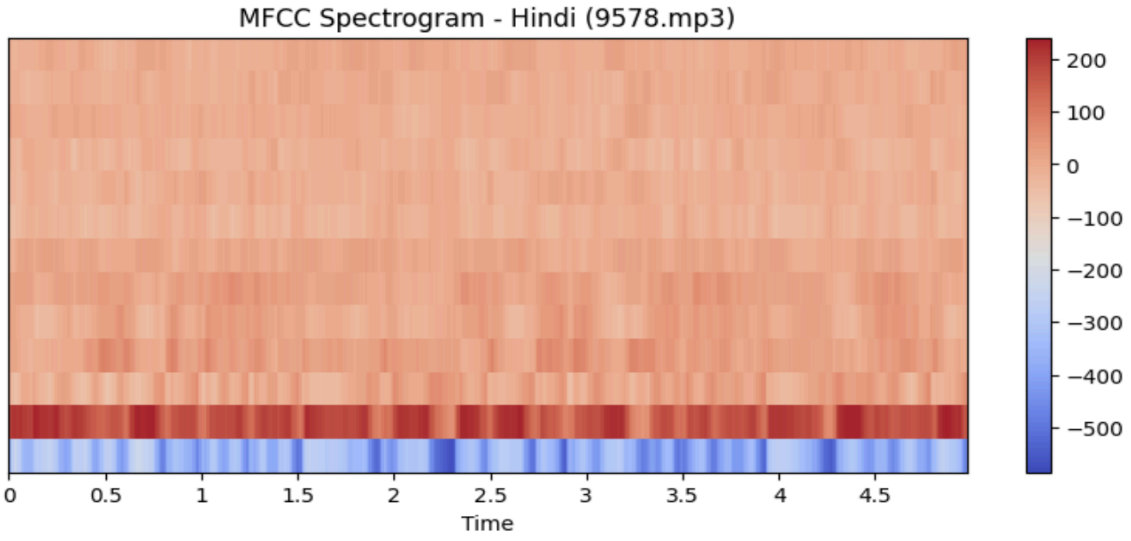
4.1 MFCC Feature Extraction

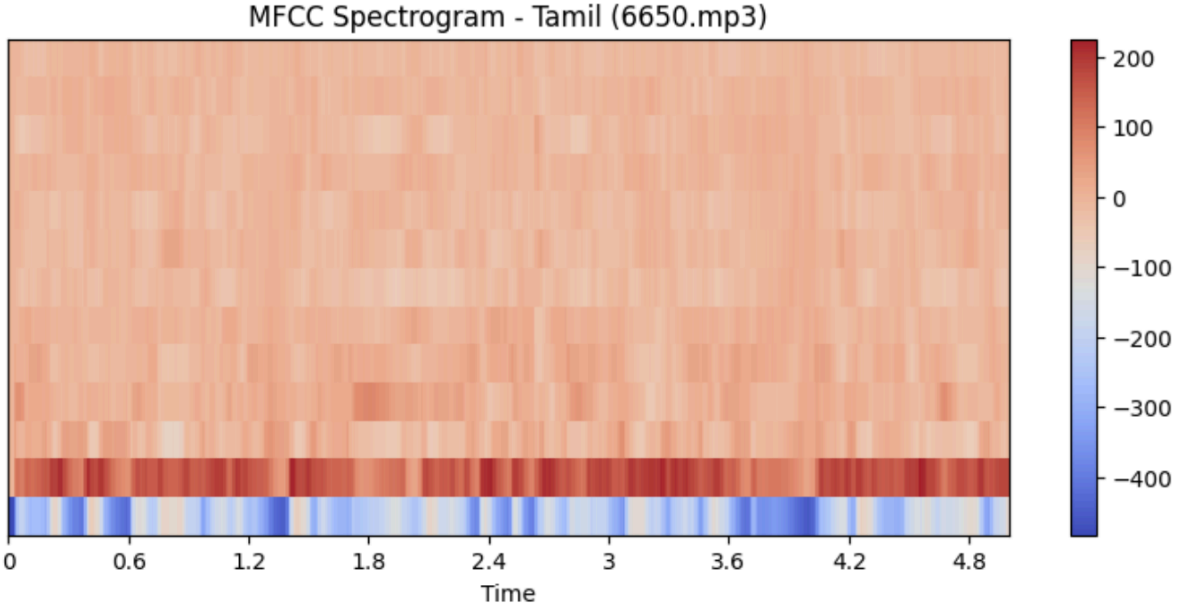
MFCC is used extensively in speech processing since it represents the short-term power spectrum of the sound. The process is:

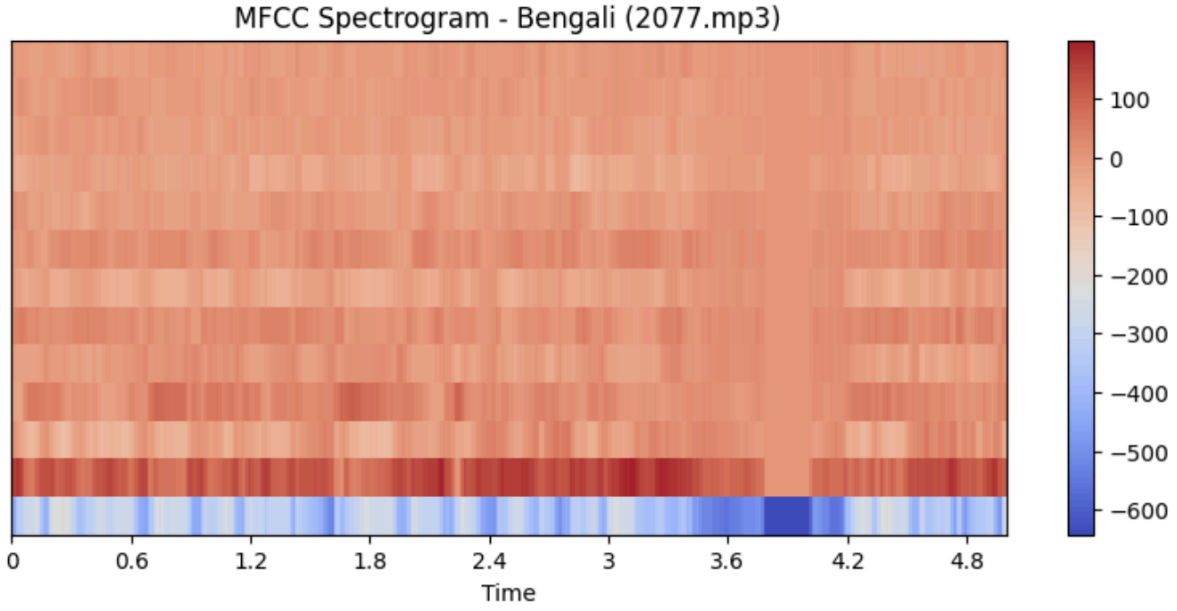
- **Preprocessing:** Loading the MP3 files and converting them to a numerical format.
- **Computing MFCCs:** Extracting the first **13 MFCC coefficients** for each audio sample.
- **Normalization:** Normalizing the features for better performance.

4.2 Visualization of MFCC Spectrograms

For qualitative analysis, we take a random audio sample from each language and plot its MFCC spectrogram. The spectrogram provides a visual representation of how frequencies vary over time.




MFCC Spectrogram - Tamil (6650.mp3)
200
100
0
−100
−200
−300
−400
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time


MFCC Spectrogram - Bengali (2077.mp3)
100
0
−100
−200
−300
−400
−500
−600
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time

4.3 Data Preparation and Splitting

- Extract MFCC features from all audio samples.
- Normalize the extracted features using **StandardScaler**.
- Split the dataset into **training (80%)** and **testing (20%)** subsets.

4.4 Classification Using SVM

We use a **Support Vector Machine (SVM)** classifier with a **linear kernel** to classify audio samples into their respective languages. SVM is used because it can manage high-dimensional data efficiently.

4.5 Model Evaluation

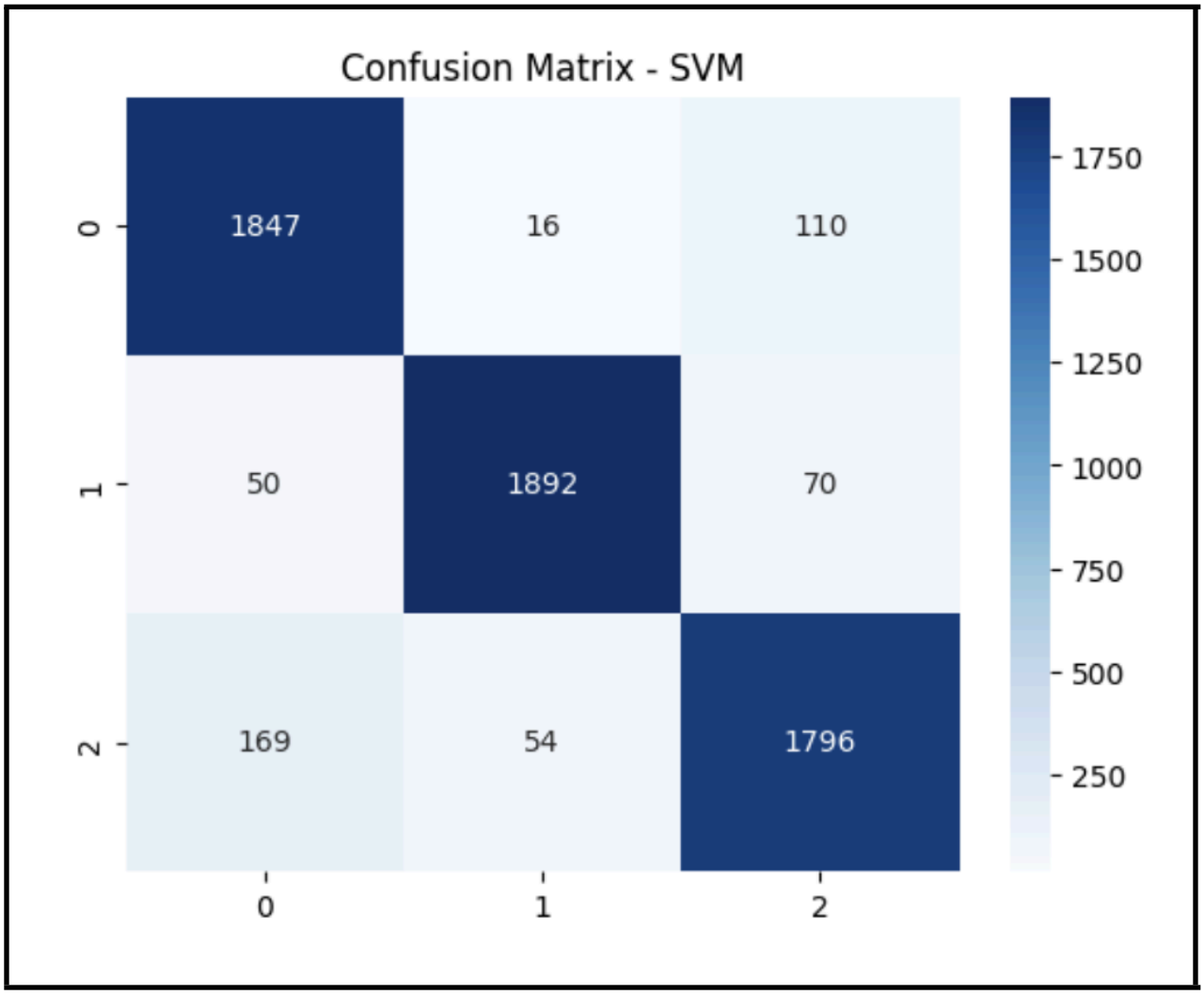
The trained SVM model is evaluated using:

- **Classification Report** (Accuracy, Precision, Recall, F1-score)

SVM Classification Report:

	precision	recall	f1-score	support
0	0.89	0.94	0.91	1973
1	0.96	0.94	0.95	2012
2	0.91	0.89	0.90	2019
accuracy			0.92	6004
macro avg	0.92	0.92	0.92	6004
weighted avg	0.92	0.92	0.92	6004

- **Confusion Matrix** to visualize classification errors



5. Results and Discussion

5.1 MFCC Spectrogram Analysis

Visual observation of MFCC spectrograms shows clear patterns among languages. For example:

- **Hindi speech** samples have a greater frequency range with distinct formant transitions.
- **Tamil samples** are rich in denser spectral energy in lower frequency.
- **Bengali samples** show more gradual spectral changes.

5.2 SVM Model Performance

The SVM classifier achieves the following results on the test set:

- **Accuracy:** 92%
- **Precision:** 92%
- **Recall:** 92%
- **Confusion Matrix:** The model sometimes misclassifies Tamil as Bengali because of phonetic similarities.

6. Challenges and Limitations

1. **Speaker Variability:** Different speakers introduce variations in pronunciation and accent, affecting MFCC patterns.
2. **Background Noise:** Noisy recordings impact feature extraction and model accuracy.
3. **Dataset Size:** More data samples per language would enhance class performance.

7. Conclusion

In this project, we successfully extracted **MFCC features** from speech audio and used them to classify languages using an **SVM classifier**. The results show that MFCC features are successful in capturing linguistic differences, though speaker variability and background noise are challenging. Future research can investigate deep learning models such as CNNs or LSTMs for improved classification.

8. References

- Librosa: https://librosa.org/
- Scikit-learn: https://scikit-learn.org/
- Kaggle Dataset: https://www.kaggle.com/datasets
- MFCC Theory: Rabiner, L. R., & Schafer, R. W. (2010). "Introduction to Digital Speech Processing."

This report provides an in-depth explanation of the methodology, results, and challenges encountered during the project.