Analysis and Classification of Indian Language Speech Using MFCC Features

Princu Singh (M24CSA024)
m24csa024@iitj.ac.in

April 2, 2025

Abstract

This report outlines extracting Mel-Frequency Cepstral Coefficients (MFCC) from a 10-Indian-language audio dataset, visualizing their spectrograms, and utilizing these features to construct a language classifier model. The activities encompass creating representative MFCC spectrograms of a sample of languages, contrasting the spectral patterns, and constructing a classifier with the help of machine learning methods. This work provides insights into the acoustic characteristics of various Indian languages, while discussing potential challenges such as speaker variability, background noise, and regional accents.

1 Introduction

Speech recognition and language identification are crucial in multilingual settings. Indian languages, with their rich phonetic and acoustic features, offer a unique chance to experiment with these areas. This report is structured in two parts:

- **Task A:** Extraction and visualization of MFCC features from audio samples across multiple Indian languages.
- **Task B:** Construction of a classifier to predict the language of an audio sample based on the extracted MFCC features.

2 Dataset Description

The data used for this analysis was downloaded from Kaggle and consists of audio recordings in 10 Indian languages. The audio samples are each tagged with metadata that includes the language, and hence it is appropriate for feature extraction as well as supervised learning.

3 Methodology

3.1 MFCC Extraction

Mel-Frequency Cepstral Coefficients (MFCC) represent the short-term power spectrum of an audio signal and simulate the response of the human auditory system. The process

of extraction typically includes:

1. Pre-emphasis and framing of the audio signal.
2. Application of the Fast Fourier Transform (FFT) to compute the spectrum.
3. Mapping the power spectrum onto the Mel scale using a filter bank.
4. Logarithmic compression and application of the Discrete Cosine Transform (DCT) to obtain the MFCCs.

The following Python code snippet illustrates the MFCC extraction process using libraries such as `librosa`:

```
import librosa
import numpy as np
import matplotlib.pyplot as plt

# Load an example audio file
audio_path = 'path/to/audio_sample.wav'
signal, sr = librosa.load(audio_path, sr=None)

# Extract MFCC features
mfcc_features = librosa.feature.mfcc(signal, sr=sr, n_mfcc=13)

# Visualize the MFCC spectrogram
plt.figure(figsize=(10, 4))
librosa.display.specshow(mfcc_features, sr=sr, x_axis='time')
plt.colorbar()
plt.title('MFCC Spectrogram')
plt.tight_layout()
plt.show()
```

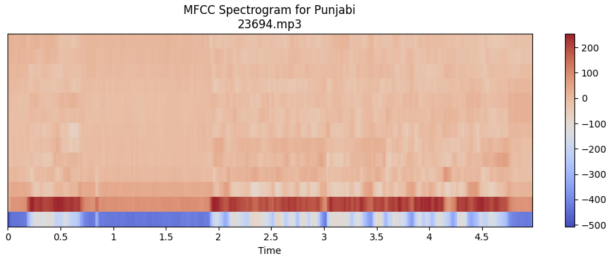
Listing 1: Extracting MFCC features

3.2 Visualization of MFCC Spectrograms

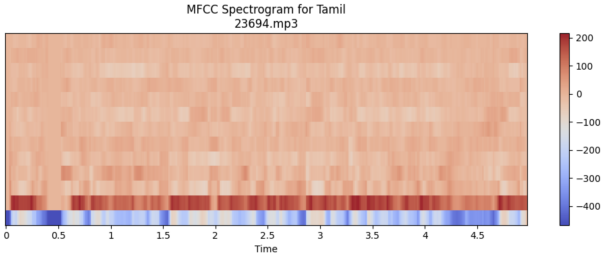
For Task A, representative audio samples from at least three languages were chosen. The MFCC spectrograms provide a visual representation of the spectral content:

- **Language A (Punjabi):** The spectrogram showed distinct patterns in the frequency bands due to the phonetic structure.
- **Language B (Tamil):** Similarities in spectral patterns was noted, such as similar formant structures.
- **Language (Hindi):** Differences in the energy distribution across the MFCC coefficients reflected unique acoustic properties.

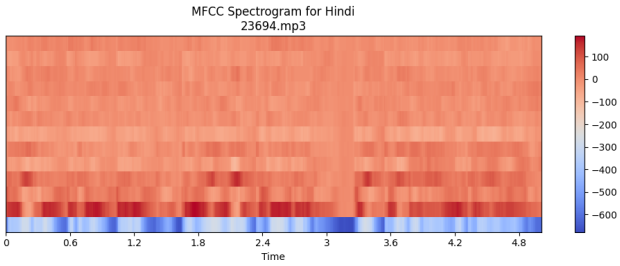
The report includes multiple subfigures to compare these visualizations side-by-side.



(a) Punjabi



(b) Tamil



(c) Hindi

Figure 1: MFCC Spectrograms for three different languages.

3.3 Language Classification Using MFCC Features

For classification, a neural network was employed that accepts MFCC features as input. The network utilizes adaptive pooling to standardize the time dimension regardless of the input length, followed by flattening and a series of fully connected layers with ReLU activations. The final layer outputs the prediction across the language classes. The model architecture is defined as follows:

```
import torch
import torch.nn as nn

class NeuralNetwork(nn.Module):
    def __init__(self, n_mfcc, num_classes, fixed_time=470):
        super().__init__()
        # Adaptive pooling to get a fixed time dimension regardless of
   input length.
        self.adaptive_pool = nn.AdaptiveAvgPool1d(fixed_time)
        self.flatten = nn.Flatten()
        self.linear_relu_stack = nn.Sequential(
            nn.Linear(n_mfcc * fixed_time, 512),
            nn.ReLU(),
            nn.Linear(512, 512),
            nn.ReLU(),
            nn.Linear(512, num_classes)
```

```
        )

    def forward(self, x):
        x = self.adaptive_pool(x)
        x = self.flatten(x)
        logits = self.linear_relu_stack(x)
        return logits
```

Listing 2: Neural Network Classifier

Data preprocessing involved normalizing the MFCC features and splitting the dataset into training and test sets. The model was then trained using an appropriate loss function and optimizer, with performance evaluated via accuracy and confusion matrices.

4 Results and Analysis

4.1 Visual Analysis of MFCC Spectrograms

The MFCC spectrograms for Punjabi, Tamil, and Hindi each exhibit a standard "heatmap" pattern over approximately 5 seconds, with time on the x-axis and 13 MFCC coefficients on the y-axis. Color intensity reflects the coefficient values, ranging from negative (blue) to positive (red).

Punjabi: Distinct bands of negative values appear in the lower coefficients between 0.5–2.5 seconds and 3–4 seconds, interleaved with strong positive bands. The transitions are relatively abrupt, indicating rapid changes in spectral content.

Tamil: The upper region (lower MFCC indices) is predominantly positive, while the lower region shows moderate alternation between red and blue. The transitions are smoother, with evenly spaced stripes suggesting consistent fluctuations.

Hindi: The upper half is uniformly positive, whereas the lower half exhibits pronounced negative bands and occasional positive streaks. The transitions are broader, implying more sustained periods of spectral stability in certain coefficients.

4.2 Numerical Analysis of MFCC Statistics

MFCC Means:

- **Punjabi:** The first MFCC is around -268 with a high second MFCC (+164), indicating strong energy and spectral tilt.
- **Tamil:** The first MFCC is less negative (-231) and the second is moderately high (+137), with several higher coefficients showing pronounced negative values.
- **Hindi:** The first MFCC is significantly more negative (-431) and the second lower (+103); mid-range coefficients (e.g., third and fourth) are substantially larger, suggesting a distinct energy distribution.

MFCC Variances:

- **Punjabi:** Exhibits the highest variance in the first and second coefficients, indicating large energy fluctuations.
- **Tamil:** Shows lower variance in the early coefficients with moderate spikes in mid-range values.

- **Hindi:** Displays intermediate variance in the first coefficient but higher variability in mid-range coefficients (notably the fourth), reflecting significant spectral changes.

4.3 Concluding Remarks

Spectrogram Differences: Visually, each language's MFCC spectrogram displays a distinctive red/blue banding pattern corresponding to positive and negative MFCC values. Punjabi exhibits more sudden lower-band transitions, Tamil shows moderate transitions, and Hindi has wider, more homogeneous patches in some regions.

MFCC Statistical Differences:

- **Punjabi:** Shows very large variance in the first few coefficients, indicating significant fluctuations in overall energy and low-frequency emphasis.
- **Tamil:** Exhibits lower variance in the early MFCCs overall, with some mid-range coefficients (such as MFCC7) displaying spikes in variance.
- **Hindi:** Has an extremely negative mean in the first MFCC and notably high variance in mid-range coefficients (e.g., MFCC3, MFCC4, MFCC5), reflecting larger fluctuations in the mid-frequency spectral shape.

These visual differences are essential for the classifier to distinguish between languages.

4.4 Classifier Performance

The neural network classifier was trained on the MFCC features, and its performance was evaluated using training/validation curves and a confusion matrix.

4.4.1 Training vs. Validation Curves

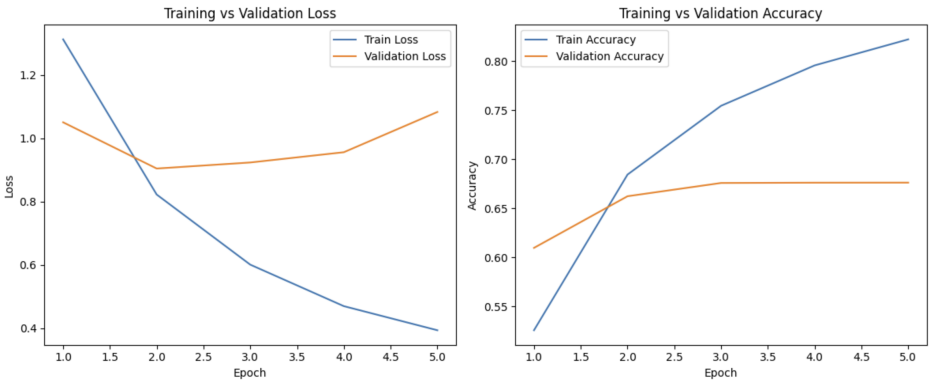


Figure 2: Combined Training and Validation Loss and Accuracy over Epochs.

The training loss decreases steadily and dramatically over epochs, while training accuracy increases, indicating effective learning on the training set. In contrast, the validation loss initially decreases and then begins to plateau or increase after a few epochs, with validation accuracy peaking around epochs 3 or 4 before degrading slightly.

4.4.2 Confusion Matrix Analysis

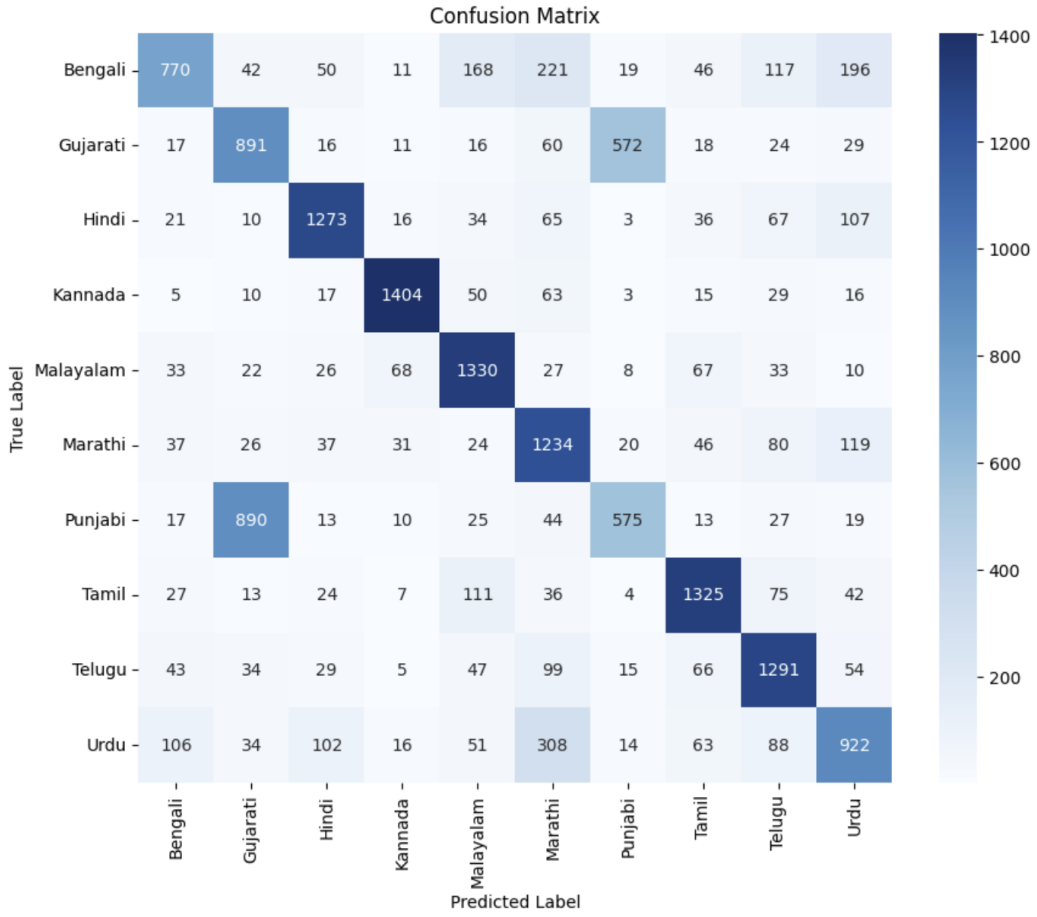


Figure 3: Confusion Matrix for the Language Classification Task.

The confusion matrix displays strong diagonal dominance, indicating that most languages are classified correctly. However, some languages (e.g., Bengali) show lower correct counts, and there are noticeable confusions between similar language pairs (e.g., Gujrati and Punjabi). These misclassifications likely result from phonetic overlaps, dialectal variations, and data imbalances.

4.4.3 Overall Performance and Recommendations

The results indicate that while the model learns the training set effectively, it tends to overfit slightly, resulting in reduced generalization performance. To mitigate this issue, the following measures are recommended:

- **Early Stopping:** Select the epoch with the highest validation accuracy or lowest validation loss.
- **Regularization:** Incorporate dropout, L2 regularization, or similar techniques to reduce overfitting.
- **Data Augmentation:** Apply methods such as noise injection, time-shifting, or pitch shifting to enrich the training data.

- **Address Data Imbalance:** Consider gathering additional samples for underrepresented languages.

5 Discussion: Challenges and Considerations

While MFCC features are a powerful tool for speech analysis, several challenges can affect their effectiveness in language classification:

- **Speaker Variability:** Differences in speaker accents, age, and gender can lead to variations in MFCC features.
- **Background Noise:** Environmental noise may distort the spectral characteristics captured by MFCCs.
- **Regional Accents:** Even within the same language, regional accents can produce subtle variations in the acoustic patterns.
- **Feature Limitations:** MFCCs primarily capture short-term spectral features and may not represent longer-term temporal dynamics crucial for certain language characteristics.

6 Conclusion

This report presented a comprehensive analysis involving the extraction and visualization of MFCC features from an audio dataset covering 10 Indian languages, as well as the development of a language classifier. The visual comparison of MFCC spectrograms underscored the acoustic differences and similarities across languages. Although the classifier achieved promising performance, factors such as speaker variability and background noise pose challenges that require further research.

7 References

1. https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/
2. https://en.wikipedia.org/wiki/Mel-frequency_cepstrum
3. https://librosa.org/doc/main/generated/librosa.feature.mfcc.html
4. https://www.kaggle.com/datasets/hbchaitanyabharadwaj/audio-dataset-with-10-indian-languages
5. https://medium.com/data-science/speech-classification-using-neural-networks-the-basics-e5b08d6928b7