Question 2 Report: MFCC Feature Extraction and Comparative Analysis

Anuj Rajan Lalla
Roll Number: B22AI061
CSL7770: Speech Understanding

Programming Assignment 2, Question 2

GitHub Repository: Click here

1 Introduction

- **Objective:** Extract and analyze MFCC features from an audio dataset of 10 Indian languages (downloaded from Kaggle).
- **Goals:**
  - Extract MFCCs from each audio sample.
  - Visualize spectrograms for representative samples (e.g., Hindi, Gujarati, Punjabi).
  - Compare MFCC patterns to identify differences and similarities.
  - Optionally, compute mean and variance for further analysis.

2 Dataset Description

- **Source:** Audio dataset from Kaggle.
- **Content:** Massive collection of audio samples from 10 different Indian languages.
- **Languages:** Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu, and Urdu.
- **Duration:** Each audio sample is approximately 5 seconds long.

- **Origin:** Created using regional videos available on YouTube.
- **Sampling Rate:** Approximately 44 kHz, which is standard for high-quality audio.
- **Scope:** Constrained to Indian languages but can be extended in future work.

3 MFCCs and Their Usefulness

- **What are MFCCs?**
  - Represent the short-term power spectrum of an audio signal.
  - Computed by taking the log of mel-scaled filter bank energies and applying DCT.
- **Why Use MFCCs?**
  - **Spectral Envelope:** Capture the overall shape of the audio spectrum.
  - **Perceptual Relevance:** Mimic human ear sensitivity via the mel scale.
  - **Dimensionality Reduction:** Provide a compact, informative representation.
  - **Classification:** Serve as effective features for input to neural networks.
- **Practical Use in Our Project:**
  - **Feature Extraction:** MFCCs are computed for each audio sample.
  - **Visualization:** Spectrograms are plotted to compare different languages.
  - **Classification:** MFCCs are fed into a CNN for language identification.

4 MFCC Extraction Process

- **Audio Loading:**
  - Audio is loaded at its original sampling rate (approx. 44 kHz).
- **MFCC Computation:**

  - 13 coefficients are computed.
  - The 0th coefficient (overall energy) is removed.

- **Efficiency:**
  - Joblib is used for parallel processing to handle multiple files.

5 MFCC Spectrogram Visualizations for Representative Samples

To illustrate the MFCC patterns across different languages, I plotted 8 samples each from Hindi, Gujarati, and Punjabi. These spectrograms provide a clear visual snapshot of the variations in MFCC coefficients over time. Figures 1, 2, and 3 show the respective plots. These 8 samples were chosen randomly

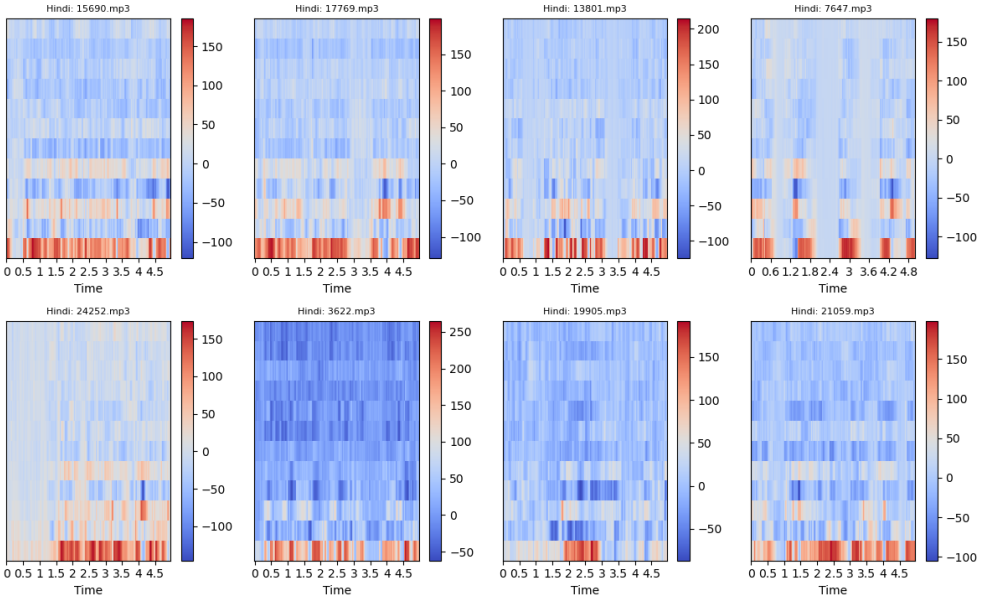

Figure 1: MFCC spectrograms for 8 representative Hindi samples.

Additional spectrograms were also generated for other languages in the dataset (such as Tamil, Telugu, Malayalam, Marathi, Kannada, Urdu, and Bengali). However I only show three languages here. The complete set of plots can be found in the accompanying Jupyter notebook.

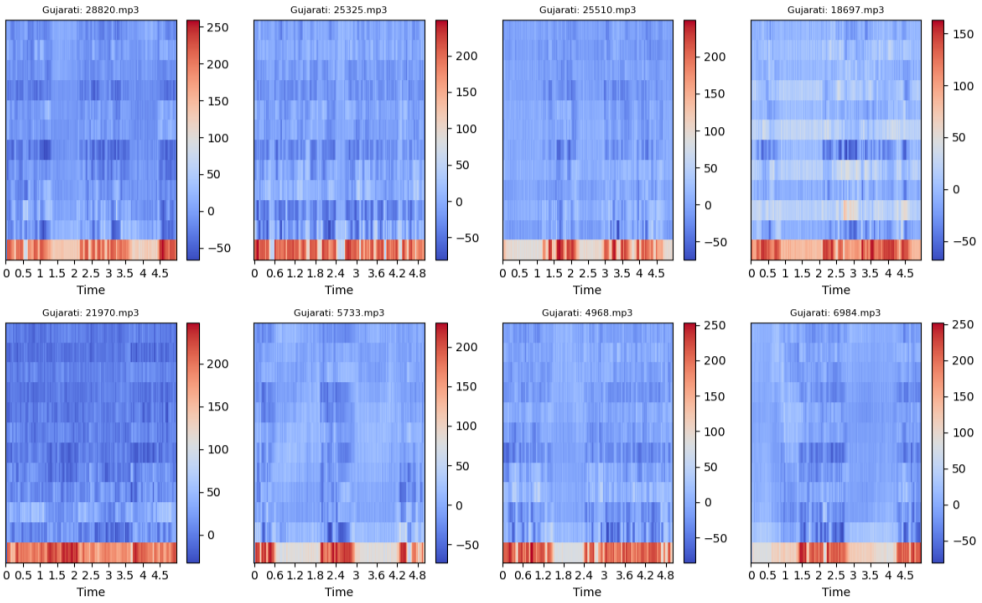

Figure 2: MFCC spectrograms for 8 representative Gujarati samples.

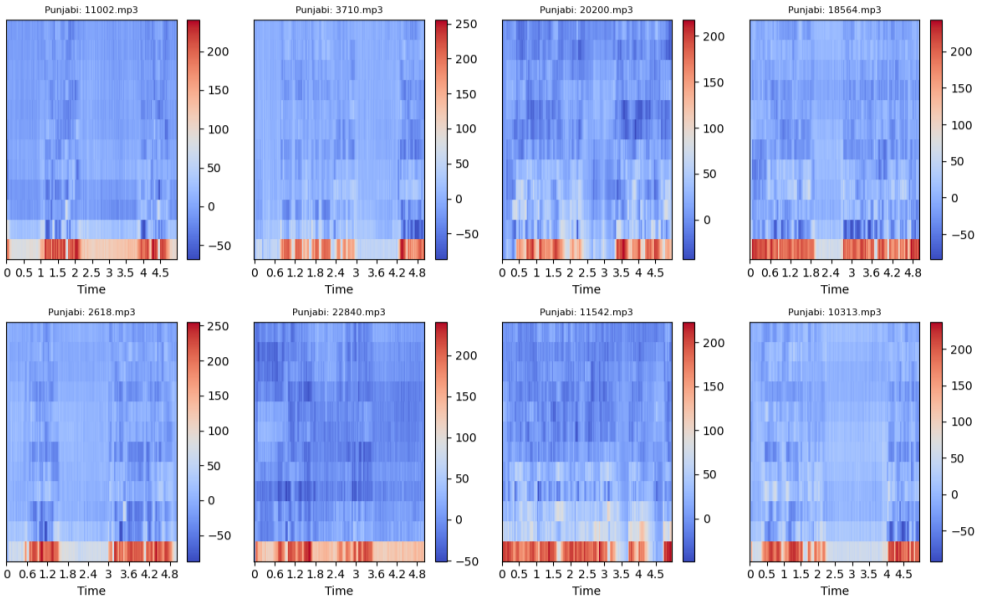

Figure 3: MFCC spectrograms for 8 representative Punjabi samples.

6 Visual Analysis and Discussion

Based on the MFCC spectrograms for Hindi, Gujarati, and Punjabi, I observed several key differences that provide insight into the acoustic characteristics of these languages.

1. Sample Plots

- **Representative Samples:** Eight representative MFCC spectrograms were plotted for each language.
- **Similar Patterns in Gujarati and Punjabi:** The spectrograms for Gujarati and Punjabi exhibit very similar patterns across the coefficients, suggesting that their overall spectral shapes and energy distributions are alike.
- **Distinct Trends in Hindi:** In contrast, the spectrograms for Hindi show noticeable differences, with lower average magnitude values across several coefficients.

2. Observations on MFCC Coefficients

- **Lower-Order Coefficients (e.g., 1st coefficient):**
  - **Broad Spectral Envelope:** These coefficients capture the overall shape of the spectrum, reflecting the basic configuration of the vocal tract.
  - **Energy Distribution:** In Gujarati and Punjabi, the higher values suggest a more pronounced energy distribution in the lower frequencies, while Hindi shows relatively lower values, indicating a less intense energy pattern.
- **Higher-Order Coefficients (e.g., 12th coefficient):**
  - **Finer Spectral Details:** These coefficients capture subtle articulatory features and fine-grained spectral characteristics.
  - **Comparison:** The more negative values in Gujarati and Punjabi imply that these languages exhibit more pronounced fine details, whereas Hindi's less negative values indicate a distinct spectral structure possibly due to different articulatory settings.

3. Implications

- **Overall Trends:** The trend across coefficients shows that Hindi generally has lower magnitude values in both broad spectral features and fine details compared to Gujarati and Punjabi.
- **Acoustic Characteristics:** This suggests that the energy distribution and vocal tract configurations in Hindi differ significantly from those in Gujarati and Punjabi.
- **Similarity Between Gujarati and Punjabi:** The similarity in the MFCC patterns of Gujarati and Punjabi supports the idea that these languages share comparable acoustic characteristics, possibly due to similar phonetic or regional influences.

4. Statistical Support

- **Numerical Reinforcement:** Detailed numerical statistics (presented later) further validate the visual trends observed in the spectrograms.
- **Confirmation of Trends:** These statistics reinforce the hypothesis that MFCC features effectively capture the distinctive spectral properties across different languages.

7 Statistical Analysis of MFCC Coefficients

Tabular Representation

Both the tables show the mean and variance of the 12 MFCC coefficients (excluding the 0th) for all 10 Indian languages. Each column corresponds to a specific coefficient index, and each language (row) has two columns: one for the mean and one for the variance.

Plots of Mean MFCC Coefficients

Figures 4 and 5 visualize the mean MFCC coefficients across all languages. The line plot shows how each language's mean values vary across the 12 coefficients, while the box plot provides a concise view of the distribution of these means.

Language	C1		C2		C3		C4		C5		C6	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
Bengali	133.27	3555.33	9.76	1502.61	23.28	786.73	6.41	560.04	6.15	352.18	-6.32	284.51
Gujarati	145.47	3176.57	-2.53	881.05	-2.86	345.05	2.78	260.58	-0.04	203.01	-12.05	308.55
Hindi	102.65	3751.15	0.32	1072.59	32.69	873.11	-7.61	764.42	17.15	391.73	-11.93	297.26
Kannada	171.57	1146.85	-65.99	946.52	-10.82	525.86	-1.46	446.12	-28.04	251.33	-10.01	157.56
Malayalam	117.02	4459.74	6.12	1074.86	15.01	786.98	8.28	545.57	9.22	286.77	-5.93	290.81
Marathi	151.48	3779.85	2.04	777.42	15.01	572.60	12.23	438.82	14.09	280.25	-7.99	232.58
Punjabi	145.71	3154.93	-2.54	876.40	-3.02	341.02	2.69	258.03	-0.09	202.03	-12.09	308.24
Tamil	145.01	2172.85	1.98	894.62	16.04	673.28	5.33	539.25	9.18	262.20	-13.95	301.58
Telugu	127.23	4271.25	-12.22	1686.51	20.57	1034.55	12.39	618.72	5.74	401.97	-13.16	407.80
Urdu	143.98	2448.46	7.78	958.26	32.36	662.88	12.51	543.09	9.65	330.58	-7.86	265.02

Table 1: Mean and Variance of MFCC Coefficients 1–6 (C1–C6) for 10 Indian Languages. (Beng: Bengali, Guj: Gujarati, Hin: Hindi, Kan: Kannada, Mal: Malayalam, Mar: Marathi, Pun: Punjabi, Tam: Tamil, Tel: Telugu, Urd: Urdu.)

Language	C7		C8		C9		C10		C11		C12	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
Bengali	2.14	181.21	1.43	164.85	-7.04	161.45	-2.89	150.79	-1.19	101.97	-3.66	103.83
Gujarati	-3.42	168.35	-3.74	120.64	-12.56	185.13	-6.91	99.38	-3.28	116.54	-5.28	142.79
Hindi	-0.20	219.02	-4.57	193.82	-6.79	164.60	-1.60	128.82	-6.78	140.01	-1.29	123.22
Kannada	1.42	147.10	-13.21	157.61	-12.70	111.28	0.43	96.79	-9.39	105.67	-5.03	94.06
Malayalam	-15.56	323.74	-11.08	180.82	-7.46	137.91	-10.85	172.16	-8.31	136.56	-6.93	91.08
Marathi	-0.95	126.53	0.80	128.33	-6.70	122.98	-8.30	117.16	-2.16	86.21	-2.73	92.93
Punjabi	-3.45	168.26	-3.72	120.00	-12.60	184.94	-6.95	99.08	-3.26	116.17	-5.30	143.00
Tamil	-10.48	227.39	-4.99	150.87	-7.84	142.73	-13.02	171.19	-4.05	96.73	-7.50	87.51
Telugu	-0.77	215.05	-6.70	196.16	-15.14	238.51	-2.83	170.37	-9.66	163.80	-6.63	125.85
Urdu	2.90	166.27	-2.57	163.48	-8.56	175.88	-2.70	122.80	-2.74	113.39	-3.21	126.70

Table 2: Mean and Variance of MFCC Coefficients 7–12 (C7–C12) for 10 Indian Languages. (Beng: Bengali, Guj: Gujarati, Hin: Hindi, Kan: Kannada, Mal: Malayalam, Mar: Marathi, Pun: Punjabi, Tam: Tamil, Tel: Telugu, Urd: Urdu.)

Observations from the Table and Plots

- **Hindi as an Outlier:**
  - The line and box plots both reveal that Hindi consistently exhibits lower mean values in the first few coefficients, with differences of approximately 20–30 units compared to Gujarati and Punjabi.
- **High First Coefficients in Kannada and Marathi:**
  - Kannada and Marathi display higher mean values in the first coefficient, suggesting a stronger energy component in the lower frequencies.
- **Similarity between Gujarati and Punjabi:**
  - Both languages show almost identical trends across most coeffi-

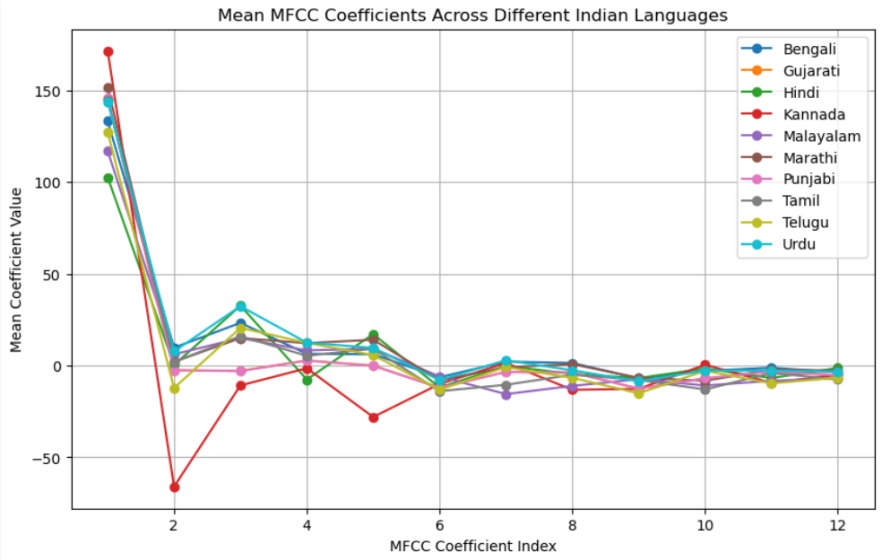

Figure 4: Line Plot of Mean MFCC Coefficients Across Indian Languages.

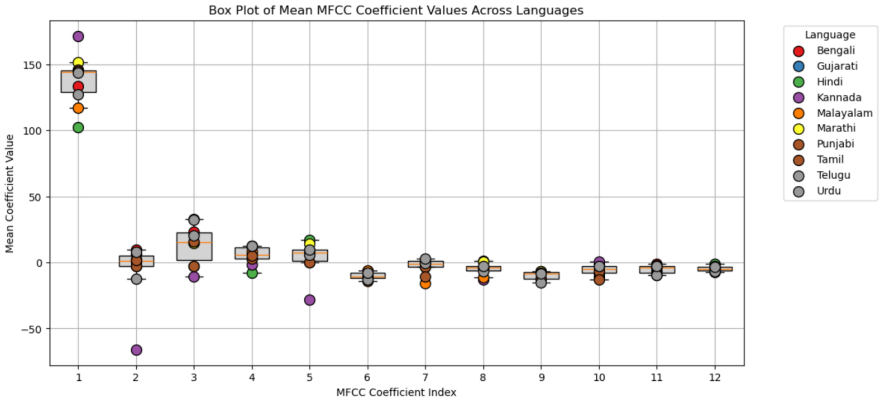

Figure 5: Box Plot of Mean MFCC Coefficient Values Across Indian Languages.

cients, reinforcing their comparable acoustic profiles.

- **Variability Across Coefficients:**
  - Some languages, such as Malayalam and Telugu, demonstrate higher variance in specific coefficients, indicating a greater dynamic range in their spectral features.

- **Implications:**
  - These numerical trends support the visual analysis, showing that MFCCs capture meaningful differences in energy distribution and spectral detail across languages.
  - The concrete numerical comparisons (e.g., the 20–30 unit difference in the first coefficient for Hindi versus Gujarati/Punjabi) provide strong evidence that these features can be leveraged for effective language classification.

8 Task B: Language Classification Using MFCC Features

In Task B, I leverage the MFCC features extracted in Task A to build a classifier that predicts the language of an audio sample. Below is a step-by-step description of the process based on my code and results.

8.1 Data Preprocessing

- **Feature Padding:** Since audio samples vary in length, I pad or truncate the MFCC arrays to a fixed shape (e.g., 12 coefficients $\times$ 200 time steps). This ensures that every sample has a uniform input size for the classifier.
- **Flattening and Label Encoding:** I convert the nested MFCC feature dictionaries into a single array of feature samples. Simultaneously, I encode language labels as integers so that the model can process them.
- **Train-Test Split and Normalization:** I split the dataset into training and testing sets (using an 80/20 split with stratification to maintain class balance). Then, I normalize the features using a global standard scaler, which helps in faster convergence during model training.

8.2 Model Selection and Architecture

- **Choice of Model:** For this task, I opted for a simple Convolutional Neural Network (CNN), which is well-suited to learning spatial patterns in MFCC spectrograms.
- **CNN Architecture:** The network consists of:

– Two convolutional layers with ReLU activations and max pooling, designed to extract hierarchical features from the MFCC input.
– Dropout layers to prevent overfitting by randomly deactivating neurons during training.
– Fully connected layers that map the extracted features to the language class predictions.

9 Model Performance and Evaluation

9.1 Training Curves

I monitored the training and validation loss over 15 epochs, as well as the corresponding accuracy curves. The loss curves (see Figure 6) demonstrate that the model converged steadily, and the accuracy curves (see Figure 7) indicate a final test accuracy of approximately 85%.

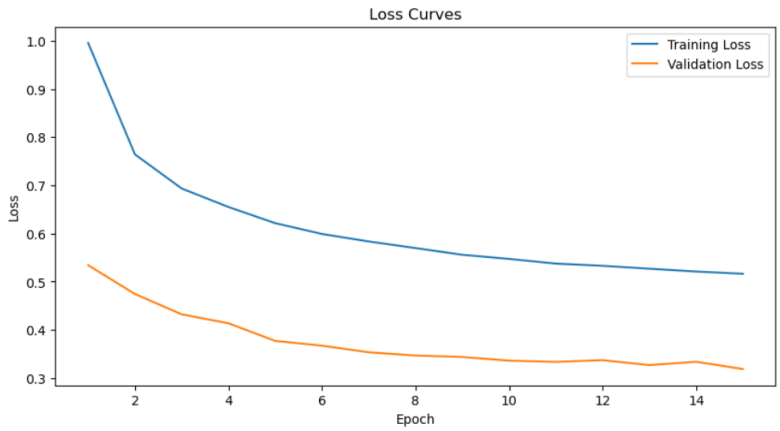

Figure 6: Training and Validation Loss Curves over 15 Epochs.

9.2 Confusion Matrix Analysis

The confusion matrix (Figure 8) reveals interesting insights. Notably, the similarities between Punjabi and Gujarati are evident, as a significant number of samples from these two languages are misclassified as one another. This supports the hypothesis that their acoustic profiles, as captured by the MFCC features, are very similar.

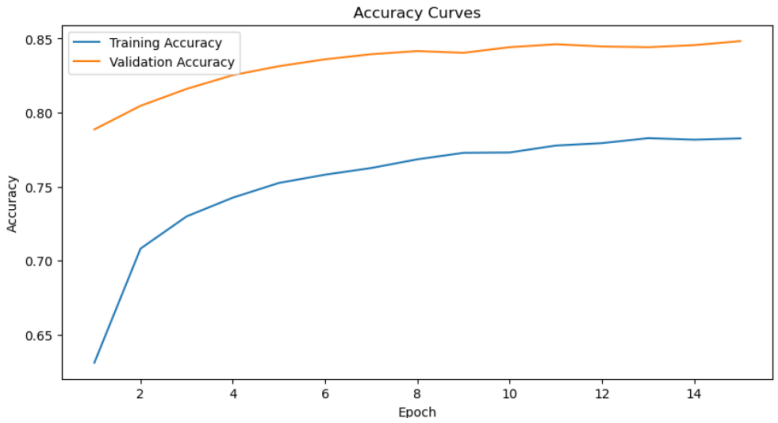

Figure 7: Training and Validation Accuracy Curves over 15 Epochs.

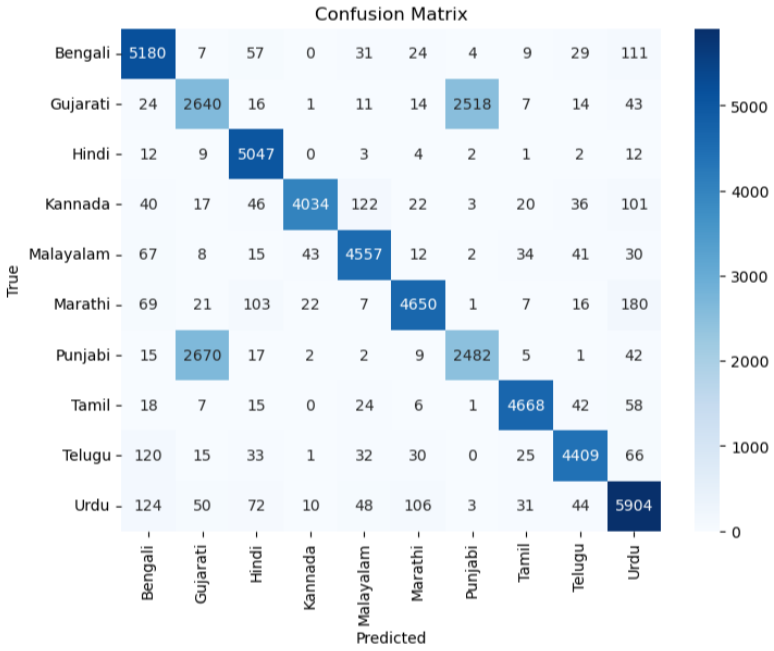

Figure 8: Confusion Matrix for the Language Classifier.

9.3 Classification Report

The detailed performance metrics for the classifier are provided below. The overall accuracy is 85%, with the following precision, recall, and f1-scores for

each language:

```
Classification Report:
              precision    recall  f1-score   support

     Bengali       0.91      0.95      0.93      5452
    Gujarati       0.48      0.50      0.49      5288
       Hindi       0.93      0.99      0.96      5092
     Kannada       0.98      0.91      0.94      4441
   Malayalam       0.94      0.95      0.94      4809
     Marathi       0.95      0.92      0.93      5076
     Punjabi       0.49      0.47      0.48      5245
       Tamil       0.97      0.96      0.97      4839
      Telugu       0.95      0.93      0.94      4731
        Urdu       0.90      0.92      0.91      6392

    accuracy                           0.85     51365
   macro avg       0.85      0.85      0.85     51365
weighted avg       0.85      0.85      0.85     51365
```

9.4 Discussion

- The final performance metrics indicate that the classifier performs well overall with an accuracy of 85%.
- Hindi, Bengali, Kannada, Tamil, Telugu, and Urdu exhibit high precision and recall, suggesting that the MFCC features capture distinctive acoustic properties for these languages effectively.
- In contrast, Gujarati and Punjabi show relatively low precision and recall. The confusion matrix supports this observation, as these two languages are often confused with each other, highlighting the similarity in their MFCC-based spectral profiles.
- These quantitative results reinforce the earlier visual and statistical analyses, confirming that while MFCC features are robust for language classification, some languages with similar acoustic characteristics (e.g., Gujarati and Punjabi) present challenges.

9.5 Discussion and Challenges

Acoustic Reflections

- The MFCC features capture essential spectral characteristics of speech—such as energy distribution and the overall spectral envelope.
- The classifier utilizes these features to differentiate between languages; for instance, it reliably identifies languages like Hindi, Bengali, Kannada, Tamil, Telugu, and Urdu that display distinct MFCC patterns.

Performance Observations

- The overall accuracy of 85% and high precision/recall for several languages confirm that MFCCs effectively capture key acoustic differences.
- Lower precision and recall for Gujarati and Punjabi, along with high confusion rates between them, indicate that their MFCC profiles are very similar.

Potential Challenges

- **Speaker Variability:**
  - Differences in individual speaker characteristics can add noise to the MFCC features.
  - For example, among the Hindi samples, one outlier exhibited significantly higher coefficient magnitudes compared to the others.
- **Accent and Regional Similarities:**
  - Similar regional accents and dialects can blur the distinctions between languages.
  - Hindi and Urdu, for instance, share overlapping phonetic features that result in comparable MFCC representations.
- **Background Noise:**
  - Variations in recording conditions and ambient noise levels can adversely affect the reliability of the extracted MFCC features.

Future Improvements

- Incorporate additional features, such as delta and delta-delta coefficients, to capture the temporal dynamics of speech.

- Experiment with more sophisticated models or ensemble techniques to better differentiate acoustically similar languages like Gujarati and Punjabi.
- Apply noise-reduction and speaker normalization techniques to mitigate the effects of background noise and speaker variability.

In summary, Task B demonstrates that MFCC features, when properly preprocessed and fed into a CNN, provide a robust basis for language classification. The experimental results—marked by an 85% overall accuracy and detailed performance metrics—underscore the effectiveness of MFCCs in capturing meaningful spectral differences among Indian languages. At the same time, challenges such as speaker variability (evident in outlier samples) and accent similarities (notably between Hindi and Urdu) highlight areas for further refinement.

References

[1] Audio dataset with 10 indian languages. https://www.kaggle.com/datasets/hbchaitanyabharadwaj/audio-dataset-with-10-indian-languages. Accessed: 2025-03-30.

[2] Python libraries for audio and machine learning. Librosa, Joblib, Matplotlib, tqdm, NumPy, Scikit-learn, and PyTorch. These libraries were used for audio processing, feature extraction, visualization, and model building. For more details, refer to their official websites: https://librosa.org, https://joblib.readthedocs.io, https://matplotlib.org, https://github.com/tqdm/tqdm, https://numpy.org, https://scikit-learn.org, and https://pytorch.org. Accessed: 2025-03-30.

[3] Audio signal processing for machine learning. https://github.com/musikalkemist/AudioSignalProcessingForML/tree/master, 2020. Includes a video explanation on MFCC extraction and practical implementation.