MFCC Feature Extraction and Comparative Analysis of Indian Languages

Assignment 2: Speech Understanding (CSL7770)

Google Colab Link

Arnav Sharma
M24CSE004

***Abstract*—This task explores the application of Mel-Frequency Cepstral Coefficients (MFCC) in analyzing and classifying Indian languages using machine learning. We begin by extracting MFCC features from an audio dataset containing ten Indian languages. A comparative analysis of MFCC spectrograms is performed to identify unique spectral patterns and statistical differences across languages. In the second phase, these extracted MFCC features are used to train a neural network classifier to predict the language of a given audio sample. The classifier achieves an accuracy of 85.89%, demonstrating the effectiveness of MFCC-based language identification. However, a notable misclassification trend is observed between Gujarati and Punjabi, with Gujarati often being incorrectly predicted as Punjabi. This highlights phonetic and spectral similarities between these languages. Additionally, we analyze performance metrics such as confusion matrices and classification reports to evaluate the model's strengths and weaknesses. The findings highlight key acoustic variations among Indian languages and the potential of deep learning for language classification tasks.**

I. Methodology

The dataset used for this study is the Audio Dataset with 10 Indian Languages from Kaggle, which contains speech recordings in ten different Indian languages: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu, and Urdu. Each language folder in the dataset consists of multiple .mp3 files, each containing spoken utterances of varying durations. These recordings serve as a rich source of phonetic and acoustic information, making them well-suited for Mel-Frequency Cepstral Coefficient (MFCC)-based analysis and classification.

The first step in processing the dataset involved loading the audio files using Torchaudio. Since some audio files contained multiple channels, they were converted to mono by averaging the channels to ensure consistency across all samples. Error handling was implemented to skip corrupted or incompatible files, ensuring that only valid audio samples were used for feature extraction.

MFCC features were extracted from each audio sample to capture the unique spectral characteristics of different languages. The MFCC transformation was performed using Torchaudio, with each sample being transformed into a 13-dimensional feature vector. The MFCC parameters included 40 Mel filters and an FFT window size of 1024. These features were then converted into NumPy arrays for further analysis. To visualize the extracted features, MFCC spectrograms were generated for a representative sample from each language. These spectrograms, plotted using Seaborn heatmaps, provided a visual comparison of the spectral patterns across languages.

To quantify the differences in MFCC features across languages, statistical analysis was performed. The mean and variance of MFCC coefficients were computed for each language to understand their distribution. The statistical analysis was presented through line plots, comparing the mean and variance of MFCC features across different languages. These visualizations helped in identifying similarities and differences among languages based on their spectral properties.

For classification, MFCC feature vectors were extracted from multiple audio samples per language. To reduce dimensionality and create a standardized representation, the mean MFCC vector was computed for each sample by averaging

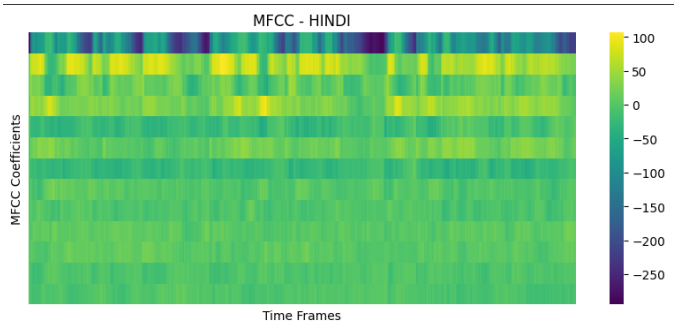

Fig. 1. Mel-Frequency Cepstral Coefficient (MFCC) Spectrogram of Hind

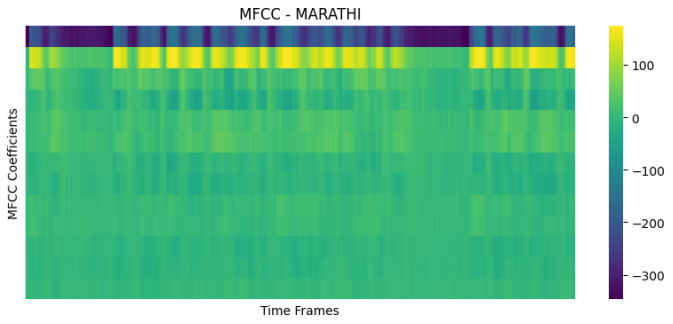

Fig. 2. Mel-Frequency Cepstral Coefficient (MFCC) Spectrogram of Marath

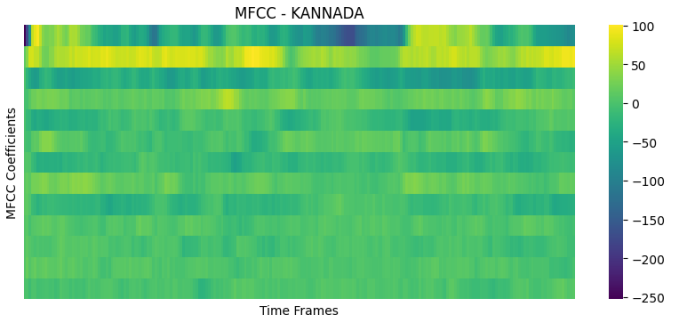

Fig. 3. Mel-Frequency Cepstral Coefficient (MFCC) Spectrogram of Kannada

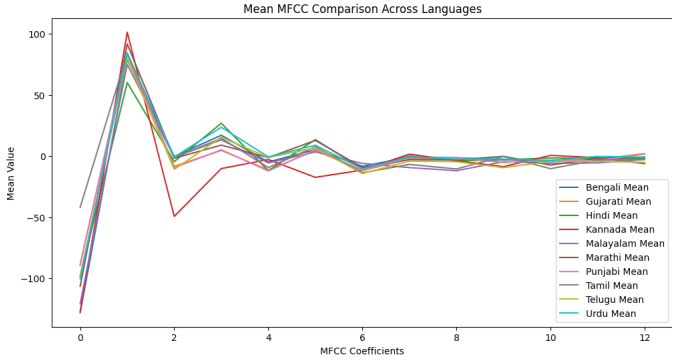

Fig. 4. Analysis of Means of Mel-Frequency Cepstral Coefficients of different languages

across time frames. Labels were assigned to each sample based on the corresponding language, and the dataset was split into 80% training data and 20% testing data to ensure proper model evaluation.

A fully connected neural network was designed using PyTorch to classify the languages based on their MFCC features. The model consisted of three fully connected layers with ReLU activation, along with batch normalization and dropout layers to prevent overfitting. The cross-entropy loss function was used for multi-class classification, and the Adam optimizer with a learning rate of 0.001 was employed to optimize the model. Training was performed for 50 epochs, using mini-batch gradient descent with a batch size of 16.

After training, the classifier was evaluated on the test dataset. The final accuracy achieved was 85.89%, indicating a strong ability to distinguish between languages based on MFCC features. However, analysis of the confusion matrix revealed that the major source of misclassification occurred between Gujarati and Punjabi. Several Gujarati samples were misclassified as Punjabi, highlighting phonetic and acoustic similarities between the two languages. To further assess the model's performance, a classification report was generated, detailing precision, recall, and F1-scores for each language. The results were visualized through bar plots comparing these metrics across different languages.

II. Results and Discussions

This section presents the performance analysis of the language classification model and examines the spectral characteristics of different languages using MFCC features. The results highlight key distinguishing factors among languages and identify challenges in classification.

A. Comparison of MFCC Spectrograms Across Languages

The Mean MFCC comparison across different languages reveals interesting insights into their spectral characteristics. The first MFCC coefficient exhibits a sharp peak for all languages, indicating the presence of strong energy in lower frequencies. However, Kannada stands out with the highest peak, suggesting that it has a more dominant low-frequency component compared to Marathi and Hindi. Marathi and Hindi, on the other hand, display a very similar pattern in the first few MFCC coefficients, hinting at close phonetic similarities between the two languages. The middle MFCC coefficients, particularly in the range of 3 to 6, show minor variations among the three languages, with Kannada maintaining a slightly flatter response, whereas Marathi and Hindi have more pronounced fluctuations. The variance of MFCC

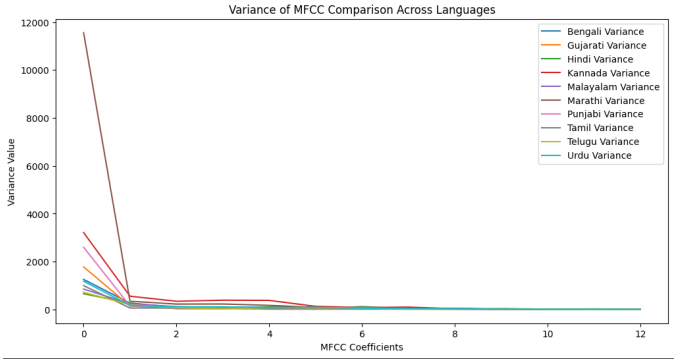

Fig. 5. Analysis of Variances of Mel-Frequency Cepstral Coefficients of different languages

coefficients further strengthens these observations. Kannada exhibits the highest variance in the first coefficient, implying that there is a significant variation in energy distribution across different samples of Kannada speech. In contrast, Marathi and Hindi demonstrate relatively lower variance, suggesting that their spectral properties remain more stable within the dataset. As the MFCC coefficient index increases, the variance diminishes for all languages, indicating that higher-order MFCC coefficients contribute less to distinguishing between them.

From these observations, it can be inferred that Kannada possesses a more distinct spectral signature compared to Marathi and Hindi, especially in the first few MFCC coefficients. This distinction could be a factor in the classification process, helping machine learning models differentiate Kannada more easily. However, the close spectral patterns between Marathi and Hindi might lead to higher misclassification rates between these two languages, as their MFCC features overlap significantly. Additionally, the high variance in Kannada's first MFCC coefficient suggests that Kannada speech samples

in the dataset cover a broader range of acoustic variations compared to Marathi and Hindi.

These findings highlight the importance of MFCC-based statistical analysis in language identification tasks. While some languages, like Kannada, may be easier to distinguish due to their unique spectral characteristics, others, such as Marathi and Hindi, may require more advanced feature engineering techniques to improve classification accuracy.

B. Classification Results and Performance Analysis

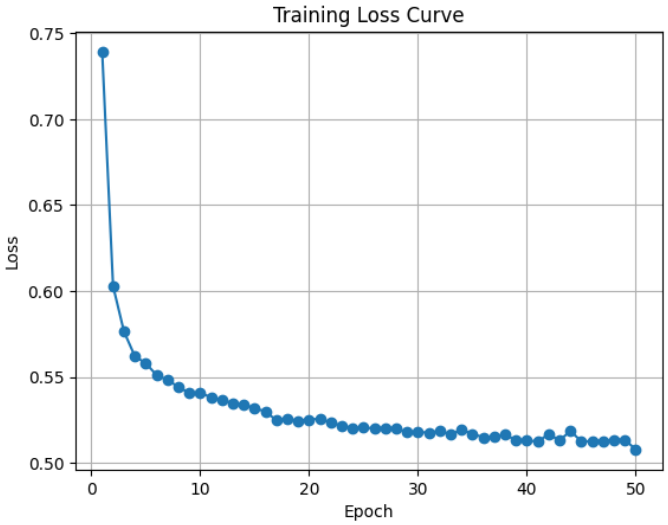

Fig. 6. Loss Curve During Model Training

The language classification model was trained on a dataset consisting of 205,460 samples, with 51,365 samples reserved for testing. The overall classification accuracy achieved was 85.89%, indicating a strong ability of the model to distinguish between different languages. However, certain misclassifications were observed, particularly between Gujarati and Punjabi, where Gujarati was often misclassified as Punjabi. This suggests that these two languages share significant phonetic similarities, making it difficult for the model to separate them effectively. The confusion matrix provides deeper insights into the model's performance. The diagonal elements of the matrix indicate the number of correctly classified samples for each language, while the off-diagonal elements represent misclassifications. Notably, the highest misclassification rates appear in classes corresponding to Gujarati and Punjabi, supporting the hypothesis that their phonetic and spectral similarities contribute to classification errors. For example, out of 5,220 actual Punjabi samples, 10 were misclassified as Hindi, 28 as Bengali, and 30 as Urdu, while Gujarati samples were often confused with Punjabi.

Another key observation from the confusion matrix is the relatively low misclassification rates between languages with distinct phonetic structures. For instance, Hindi, Kannada, and Tamil, which have different phonetic and syllabic structures, show minimal confusion with one another. This aligns with the earlier MFCC analysis, where Kannada demonstrated a

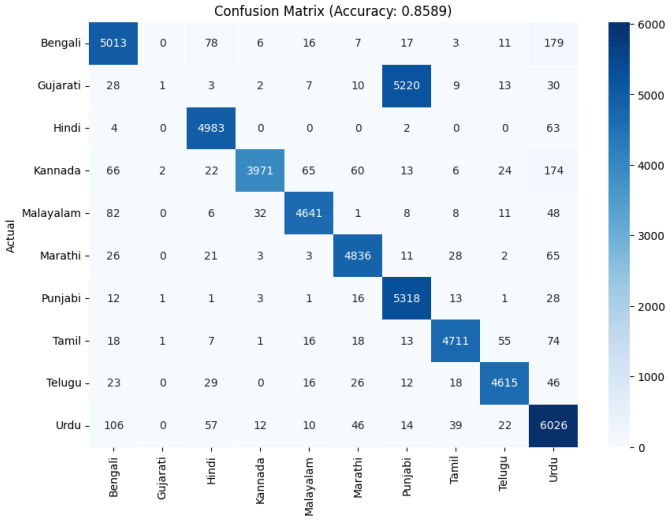

Fig. 7. Confusion Matrix of classification of 10 Indian Languages

unique spectral signature, making it easier to classify. Despite

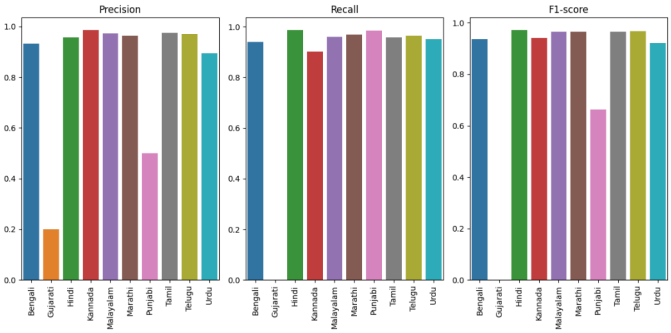

Fig. 8. Precision, Recall and F1-Score of classification of 10 Indian Languages

the strong overall accuracy, there is room for improvement, particularly in distinguishing closely related languages such as Gujarati and Punjabi. Future improvements could include augmenting the dataset with more diverse speech samples, incorporating additional acoustic features beyond MFCCs, or employing deep learning models such as transformers for enhanced feature extraction.

III. Conclusions

This study explored the effectiveness of Mel-Frequency Cepstral Coefficients (MFCC) for both the analysis and classification of Indian languages. The MFCC feature extraction process successfully captured the spectral characteristics of different languages, enabling a comparative analysis of their phonetic properties. Through visualization and statistical analysis of MFCC spectrograms, distinct acoustic patterns were identified, particularly highlighting the differences between languages like Kannada and the more similar Marathi and Hindi.

The language classification model achieved an accuracy of 85.89%, demonstrating the potential of MFCC-based features

for automatic language identification. However, challenges arose in distinguishing languages with similar phonetic and spectral traits, such as Gujarati and Punjabi, which led to notable misclassifications. Despite this, the results underscore the value of MFCC in language recognition tasks, with particular attention needed for languages that share close acoustic properties.

REFERENCES

[1] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). *librosa: Audio and Music Signal Analysis in Python*. Proceedings of the 14th Python in Science Conference. Available at: https://librosa.org/

[2] Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lax, S., and others. (2020). *Seaborn: statistical data visualization*. Journal of Open Source Software, 5(49), 2439. Available at: https://seaborn.pydata.org/

[3] Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, 9(3), 90-95. Available at: https://matplotlib.org/

[4] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köhler, J., Yang, E., and DeVito, Z. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Advances in Neural Information Processing Systems (NeurIPS). Available at: https://pytorch.org/

[5] Kaggle. (2023). *Audio Dataset with 10 Indian Languages*. Available at: https://www.kaggle.com/datasets/datrush/10-indian-languages-audio-dataset