Click here to access all the codes

REPORT

SPEECH UNDERSTANDING
CSL7770

ASSIGNMENT -2

MFCC FEATURE EXTRACTION AND COMPARATIVE ANALYSIS OF INDIAN LANGUAGES

Submitted By
Himanshu (M24CSE009)

ABSTRACT

This report presents a comprehensive study on the extraction of Mel-Frequency Cepstral Coefficients (MFCC) from audio samples of ten Indian languages and the subsequent use of these features for language classification. In the first part, MFCC spectrograms for Punjabi, Bengali, and Telugu are generated and analyzed to identify distinctive spectral characteristics. A statistical analysis of the MFCC features—focusing on mean and variance—is then conducted across all ten languages to highlight their phonetic and acoustic differences. In the second part, a neural network classifier is built using the extracted MFCC features, achieving notable accuracy with an emphasis on the challenges encountered in classifying languages such as Punjabi and Gujarati.

1 INTRODUCTION

The study of speech signals through feature extraction is critical in understanding the unique acoustic properties of languages and in developing robust language identification systems. In this report, we investigate the Mel-Frequency Cepstral Coefficients (MFCCs) as a tool for capturing the spectral characteristics of audio signals. MFCCs are widely used in speech processing because they represent the short-term power spectrum of a sound in a manner that mimics the human auditory system. Here, we focus on ten Indian languages—Punjabi, Tamil, Hindi, Bengali, Telugu, Kannada, Gujarati, Urdu, Marathi, and Malayalam—and provide a comparative analysis both in terms of the visual structure of MFCC spectrograms and statistical properties (mean and variance). Furthermore, we build a classifier using these features to automatically predict the language of an audio sample, which is crucial for applications in multilingual speech processing and language recognition systems.

2 MFCC FEATURE EXTRACTION AND VISUALIZATION

2.1 Extraction Process

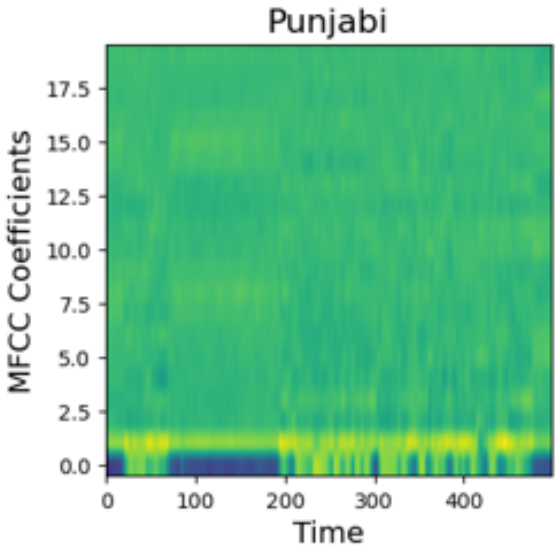
MFCC extraction begins with pre-processing, including normalization, framing, and windowing. A Fast Fourier Transform (FFT) converts each frame to the frequency domain, followed by a Mel filter bank to

emphasize perceptually relevant frequencies. Finally, the log energies undergo a Discrete Cosine Transform (DCT) to generate the MFCCs.

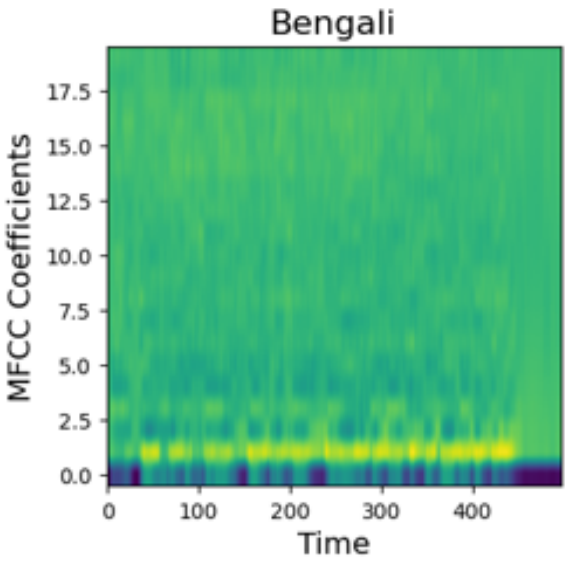
2.2 MFCC Spectrogram Visualization

To better understand the differences between languages, MFCC spectrograms were generated for a representative subset of languages—Punjabi, Bengali, and Telugu. The spectrograms provide a visual representation of how the energy in various MFCC coefficients evolves over time.

- **Punjabi:**

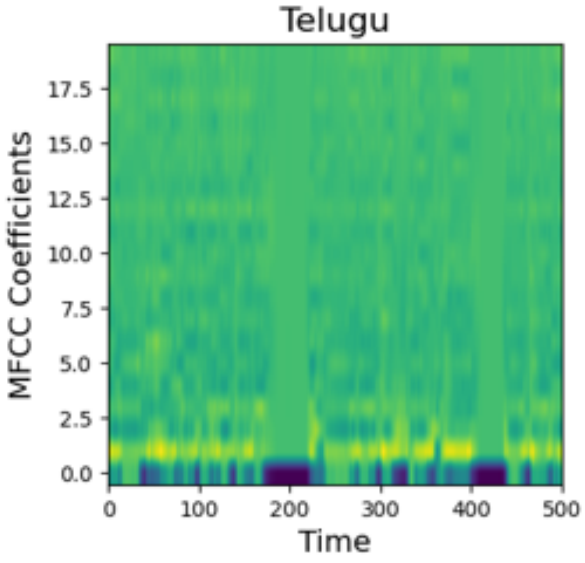


  The MFCC spectrogram of Punjabi exhibits a concentration of energy primarily in the lower coefficients (0–4), which implies a strong presence of voiced sounds and a relatively flat rhythm with less variation in pitch. The energy concentration and smooth transitions suggest that the language is characterized by steady-state vocal tract configurations.
- **Bengali:**



The Bengali spectrogram, while also showing dominant low-frequency energy, features more intricate patterns in the mid-range coefficients. This complexity is likely a result of aspirated consonants, nasalization, and other phonetic characteristics that add richness to the acoustic signal. The subtle variations across the mid-frequency range indicate a denser acoustic profile with a mixture of voiced and unvoiced segments.

- **Telugu:**



Telugu displays the most temporal variation among the three. Its spectrogram reveals clear gaps and intermittent pauses, along with a wider spread in mid-to-high frequency coefficients. These features align with Telugu's syllable-timed rhythm, where the language's structure results in dynamic pitch and formant transitions. The distinct segmentation in the spectrogram corresponds to the rhythmic alternation between vowel-rich syllables and consonantal elements.

The visual comparison of these spectrograms not only underscores the phonetic diversity of these languages but also emphasizes the potential of MFCCs to capture subtle acoustic differences that are fundamental for language identification tasks.

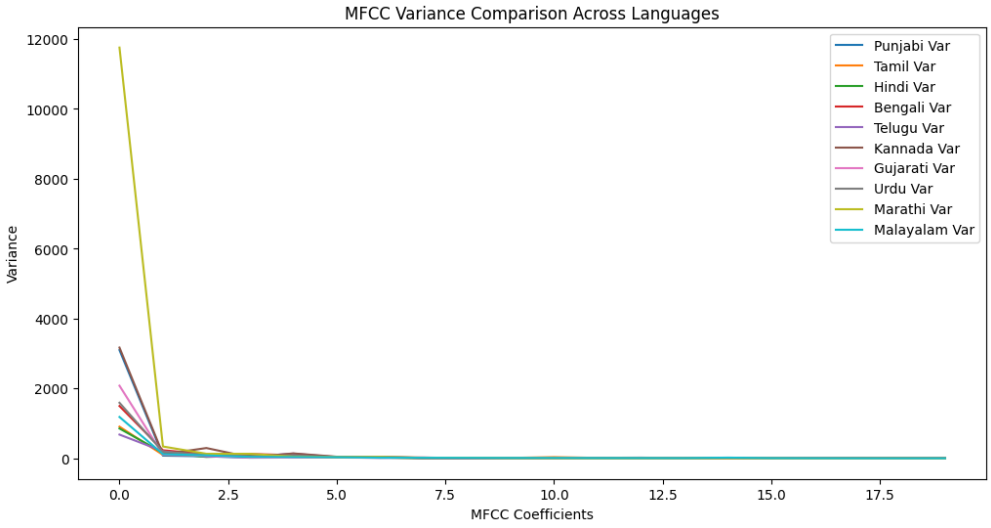
3 STATISTICAL ANALYSIS OF MFCC FEATURES

3.1 Overview

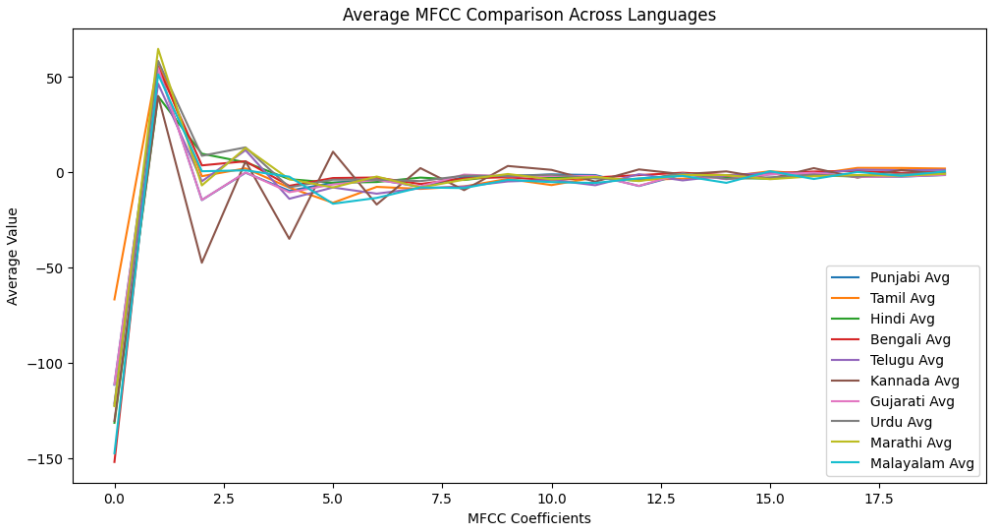
Beyond visual inspection, a quantitative analysis was performed by computing the average and variance of the MFCC coefficients for each language. This statistical analysis helps in understanding the distribution, central tendencies, and spread of the acoustic features, thereby providing a more rigorous basis for comparing the spectral characteristics across languages.

3.2 3.2. Key Findings

- **Variance Analysis:**
  The variance plot indicates that the first MFCC coefficient exhibits considerably higher variability than the rest. For instance, Marathi shows the highest variance at approximately 11,800, suggesting that its speech patterns are marked by strong spectral dynamics. In contrast, languages like Malayalam exhibit much lower variance, indicating more consistent and less fluctuating spectral features. Other languages such as Tamil, Telugu, and Kannada also display moderate to high variability, which can be linked to their distinctive phonetic constructs.



- **Average MFCC Analysis:**
  In analyzing the mean values, it is observed that the first coefficient varies widely across languages, ranging from around -150 to -50. This discrepancy is indicative of differences in the overall energy distribution in the lower frequency bands. However, for higher-order coefficients, the mean values tend to converge across languages. Notably, Marathi and Tamil stand out with higher average values in the early coefficients, while Kannada exhibits greater fluctuations in the mid-range. These results confirm that while the lower-order coefficients capture language-specific characteristics effectively, higher-order coefficients tend to encode more general spectral envelope properties.



3.3 Implications

The statistical differences in MFCC features across languages have significant implications for applications such as automated language identification and multilingual speech processing. The high variance in the first coefficient for some languages suggests that incorporating these features into a classifier could lead to better discrimination of languages that have more dynamic acoustic patterns. Moreover, the convergence of higher-order coefficients indicates a commonality in the spectral envelope structure, which could be exploited in the feature selection process for building robust classifiers.

4 LANGUAGE CLASSIFICATION USING MFCC FEATURES

4.1 Model Selection and Data Preprocessing

In the second part of the assignment, the focus shifts to utilizing the extracted MFCC features to build a classifier capable of predicting the language of an audio sample. A neural network model was chosen for this task due to its capacity to model complex, non-linear relationships in high-dimensional data.

Prior to model training, the data was preprocessed extensively. This included normalization to ensure that the MFCC features are on a comparable scale, as well as a careful train-test split to maintain the integrity of the evaluation process. These steps are critical to avoid issues such as overfitting and to ensure that the model generalizes well to unseen data.

4.2 Neural Network Architecture

The classifier architecture consisted of:

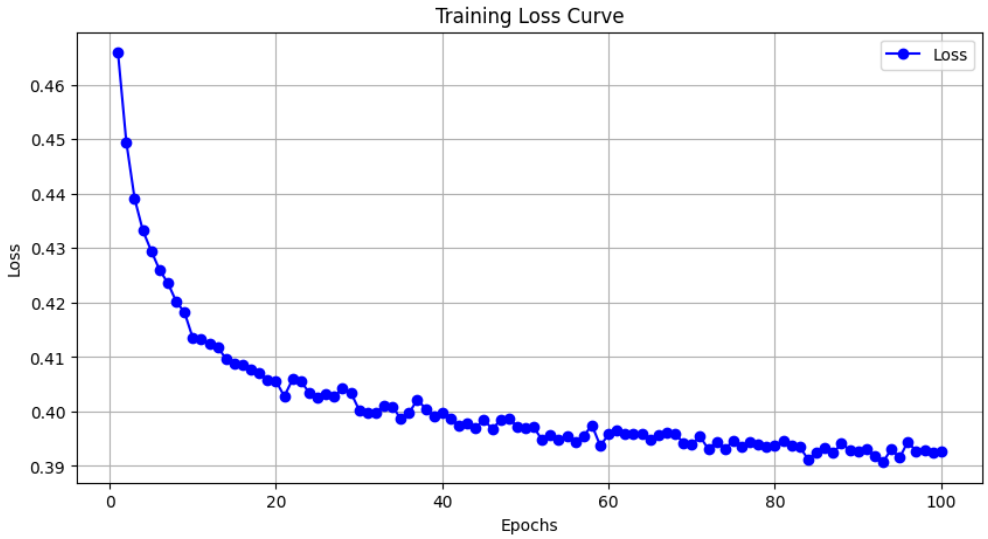
- **Input Layer:** Receives the normalized MFCC features.
- **Two Hidden Layers:** Each employs ReLU (Rectified Linear Unit) activation functions to introduce non-linearity. Dropout layers are incorporated to mitigate overfitting. These layers are designed to capture intricate patterns in the MFCC data.
- **Output Layer:** A linear layer followed by a softmax activation function, producing probability distributions across the ten language classes.

The model was trained using the Adam optimizer and cross-entropy loss function. Training was conducted over 100 epochs, during which both the training loss and test accuracy were monitored.

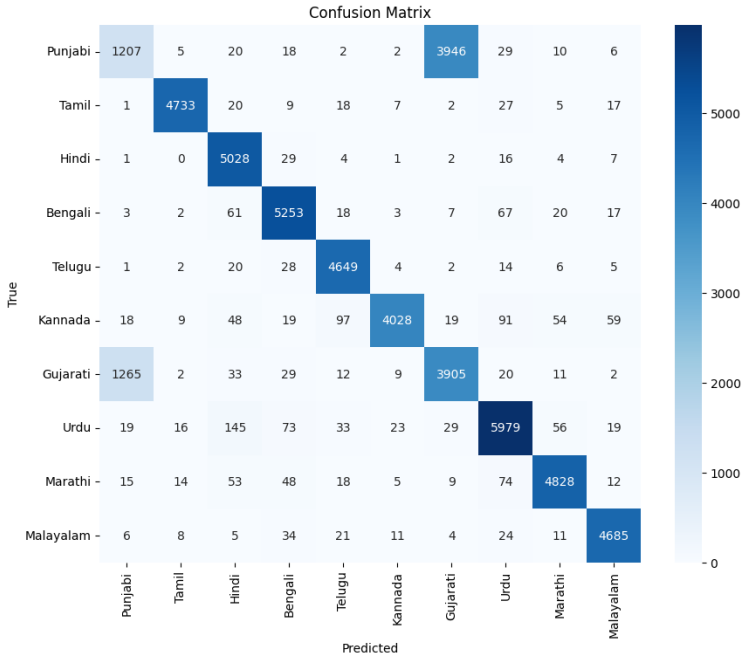
4.3 Evaluation and Results

The performance metrics obtained from the training and testing phases reveal promising results:

- **Training Loss:** The training loss converged to approximately 0.3927, indicating a good fit to the training data.



- **Test Accuracy:** The test accuracy was recorded at 86.24%, demonstrating strong generalization performance on unseen data.
- **Class-wise Performance:** The confusion matrix and classification report highlight performance differences across languages.
  - Languages such as **Tamil, Hindi, Bengali, Telugu, Kannada, Urdu, Marathi**, and **Malayalam** achieved high precision, recall, and F1-scores (generally above 0.90).
  - **Punjabi** and **Gujarati** were less accurately classified:
    - Punjabi: Precision = 0.48, Recall = 0.23
    - Gujarati: Precision = 0.49, Recall = 0.74



These discrepancies suggest that while the classifier performs well overall, there remain challenges in distinguishing languages with similar acoustic profiles or those that might have overlapping spectral features. Addressing these challenges might involve incorporating additional features, refining the network architecture, or applying advanced techniques such as data augmentation to balance the representation of all language classes.

5 CONCLUSION

This report presents a comprehensive examination of MFCC feature extraction, visual and statistical comparative analysis of Indian languages, and the subsequent development of a language classification model. The findings demonstrate that:

- **MFCC Spectrograms:** MFCC spectrograms can effectively capture distinctive acoustic characteristics across languages.
- **Statistical Measures:** Statistical measures such as mean and variance of MFCC coefficients provide valuable insights into the spectral dynamics inherent to each language.
- **Neural Network Classifier:** A neural network classifier trained on normalized MFCC features can achieve robust performance in language identification, though further refinement is needed for languages that exhibit overlapping spectral characteristics.

REFERENCES

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library*. Advances in Neural Information Processing Systems (NeurIPS).

- Snyder, S., Chen, N., Povey, D., & Khudanpur, S. (2015). *MUSAN: A Music, Speech, and Noise Corpus*. arXiv preprint arXiv:1510.08484.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

- Bharadwaj, C. (2023). *Audio Dataset with 10 Indian Languages*. Kaggle. https://www.kaggle.com/datasets/hbchaitanyabharadwaj/audio-dataset-with-10-indian-languages
  Reference for: Dataset used in this study