Statistical Analysis of MFCC Features for Indian Languages

Extraction, Analysis, and Classification
Question 2

Harishankar Nagar (D24CSA003)

Abstract

This report presents an analysis of Mel-Frequency Cepstral Coefficients (MFCC) extracted from audio samples of four Indian languages: Hindi, Malayalam, Urdu, and Bengali. The objective was to investigate whether these acoustic features can effectively differentiate between languages. MFCC features were extracted from 4,000 audio samples per language and statistical analysis was conducted to identify significant differences between languages. The findings showed that MFCC features exhibit distinct patterns across languages, with statistical significance confirmed through ANOVA and pairwise t-tests. These results suggest that MFCC features provide a robust foundation for automated language identification systems. Building upon this analysis, several machine learning models were trained and evaluated that successfully classified languages based on MFCC features, with a Support Vector Machine classifier achieving the highest accuracy of 97.7

1 Introduction

Language identification is a fundamental task in speech processing with applications ranging from multilingual speech recognition to automated customer service systems. In linguistically diverse regions like India, with over 22 officially recognized languages, automated language identification systems can significantly improve accessibility and efficiency of speech-based services.

This task investigates the effectiveness of Mel-Frequency Cepstral Coefficients (MFCC) [8] as acoustic features for distinguishing between different Indian languages. MFCCs are widely used in speech processing due to their ability to represent the short-term power spectrum of sound in a way that approximates human auditory perception.

The study focuses on four Indian languages: Hindi, Malayalam, Urdu, and Bengali. These languages represent diverse language families spoken across different regions of India, making them suitable candidates for investigating acoustic differences. Hindi and Urdu share linguistic similarities but use different writing systems, while Malayalam (Dravidian family) and Bengali (Indo-Aryan family) offer phonological contrast.

The statistical properties of MFCC features across these languages were analyzed to:

- Identify whether distinct acoustic patterns exist across languages

- Quantify the significance of these differences through statistical testing
- Establish the viability of MFCC features for language identification tasks
- Build and evaluate classification models for automated language identification

2 Data and Methodology

2.1 Dataset

The analysis used the ”Language Detection Dataset” from Kaggle [4], which contains audio samples from 10 Indian languages. For this study, four languages were selected:

- Hindi: 25,462 audio samples
- Malayalam: 24,044 audio samples
- Urdu: 31,960 audio samples
- Bengali: 27,258 audio samples

To ensure balanced representation, 4,000 audio samples were randomly selected from each language for feature extraction and analysis.

2.2 MFCC Feature Extraction

Mel-Frequency Cepstral Coefficients were extracted using the following process:

1. Audio signals were loaded with a sampling rate of 22,050 Hz
2. 13 MFCC coefficients were extracted from each audio sample using the Librosa Python library [5]
3. MFCC features were standardized to a uniform length of 215 frames through truncation or zero-padding

The code for feature extraction was implemented in two Python scripts:

- `mfcc_extractor.py`: Core functionality for MFCC extraction
- `mfcc_analysis.py`: Analysis and visualization of the extracted features

2.3 Statistical Analysis Methods

To quantify differences between languages, the following statistical analyses were performed:

- Calculation of mean and standard deviation of MFCC coefficients for each language
- One-way ANOVA to test for statistically significant differences in MFCC coefficients across languages
- Pairwise t-tests between languages to identify which language pairs exhibit significant differences

Additionally, MFCC spectrograms were generated for visual comparison of acoustic patterns across languages.

3 MFCC Analysis Results

3.1 Visual Analysis of MFCC Spectrograms

The MFCC spectrograms revealed distinctive patterns across the four languages. Figure 1 presents a grid of representative MFCC spectrograms for three samples from each language.

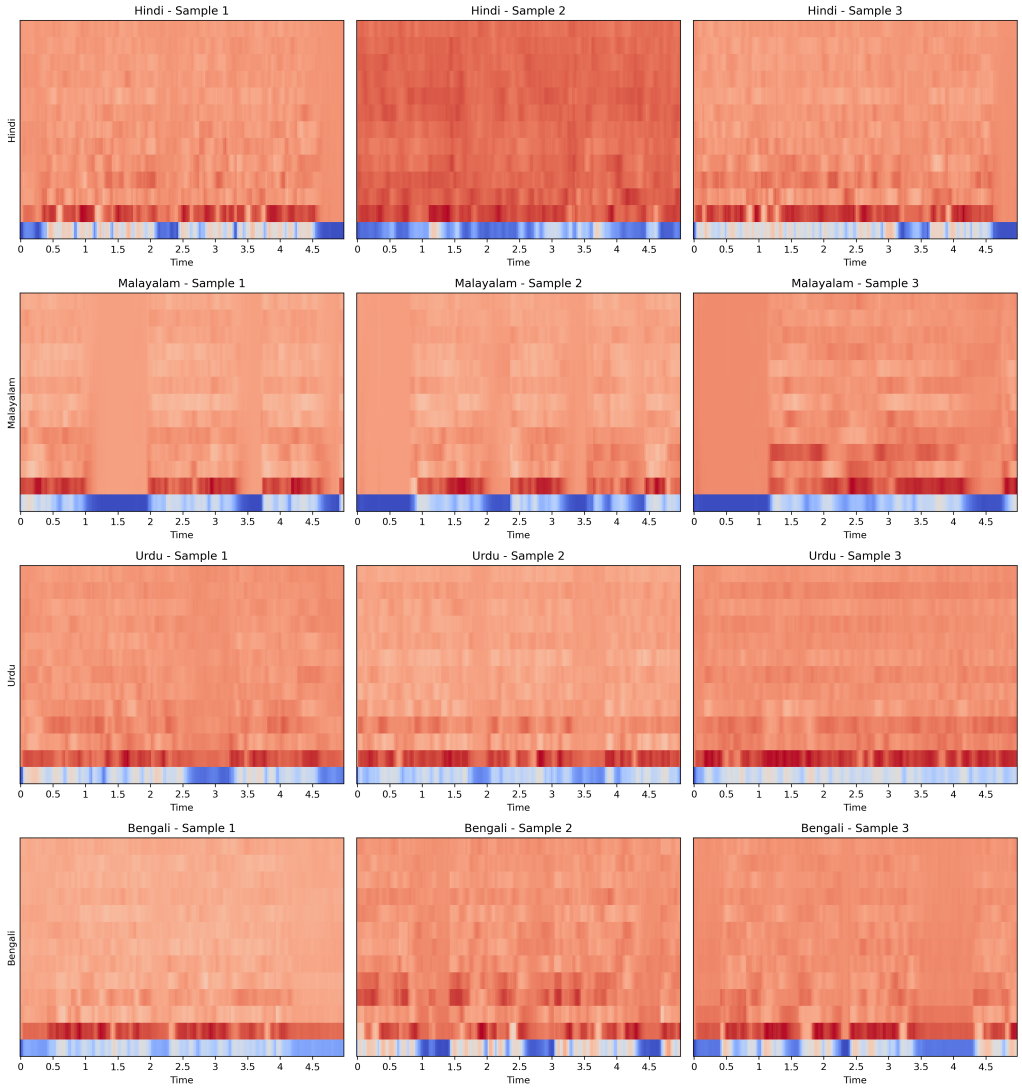

Figure 1: MFCC spectrograms for Hindi, Malayalam, Urdu, and Bengali (3 samples each)

Visual observation of the spectrograms revealed several patterns:

- Hindi samples typically showed strong energy concentration in the lower MFCC coefficients with distinct temporal variations
- Malayalam spectrograms exhibited more frequent transitions between high and low coefficient values
- Urdu displayed a more uniform pattern across the time axis with characteristic bands in the lower coefficients

- Bengali showed moderate variability with energy distribution across a wider range of coefficients

These visual differences suggest that the languages possess distinct acoustic characteristics that were captured by the MFCC representation.

3.2 Statistical Distribution of MFCC Coefficients

The mean values of MFCC coefficients across languages are visualized in Figures 2 and 3, while Figure 4 shows the standard deviations.

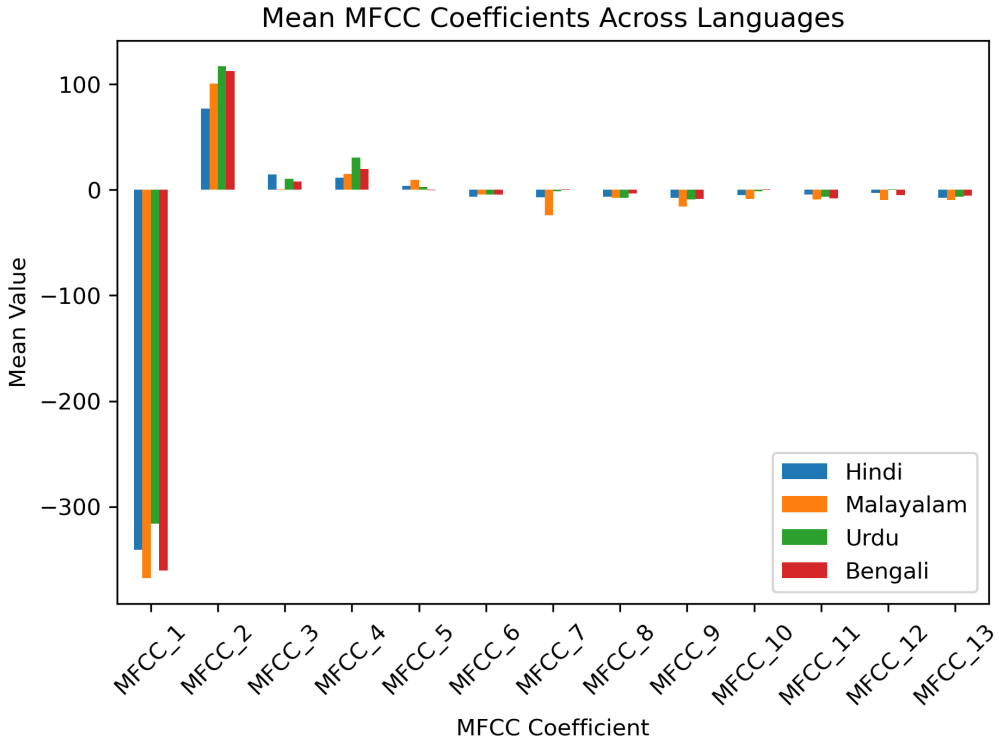

Figure 2: Mean MFCC coefficient values for each language

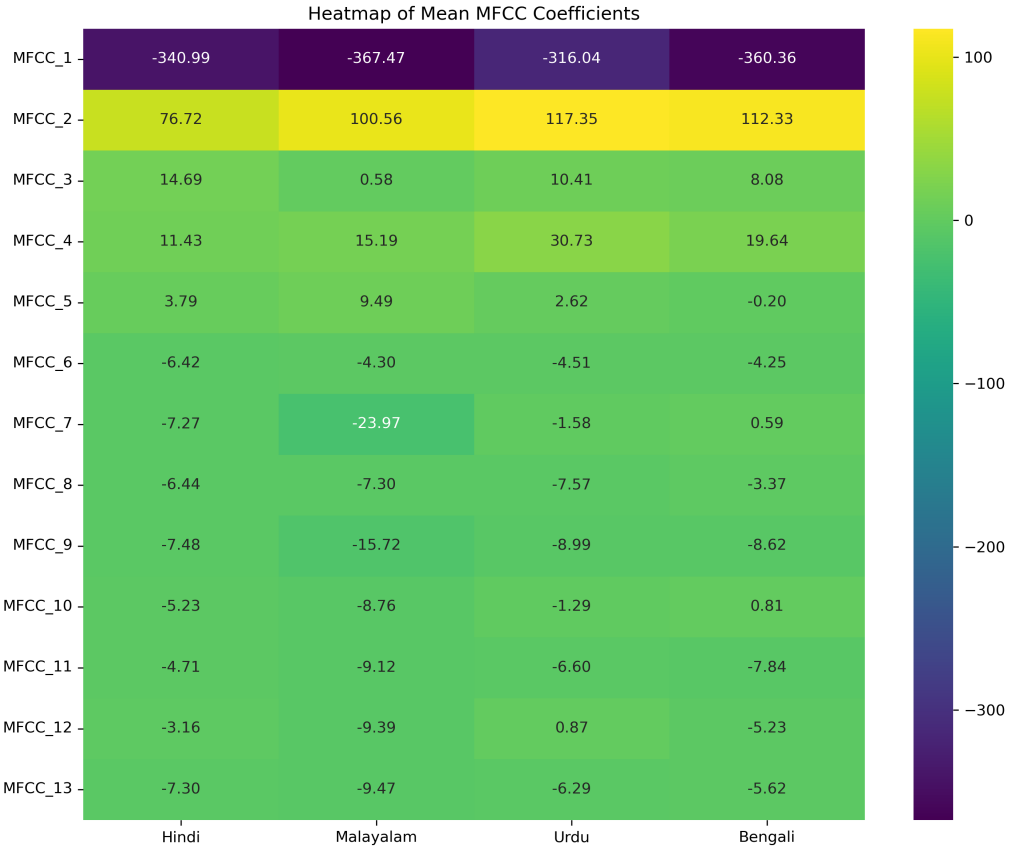

Figure 3: Heatmap of mean MFCC coefficients across languages

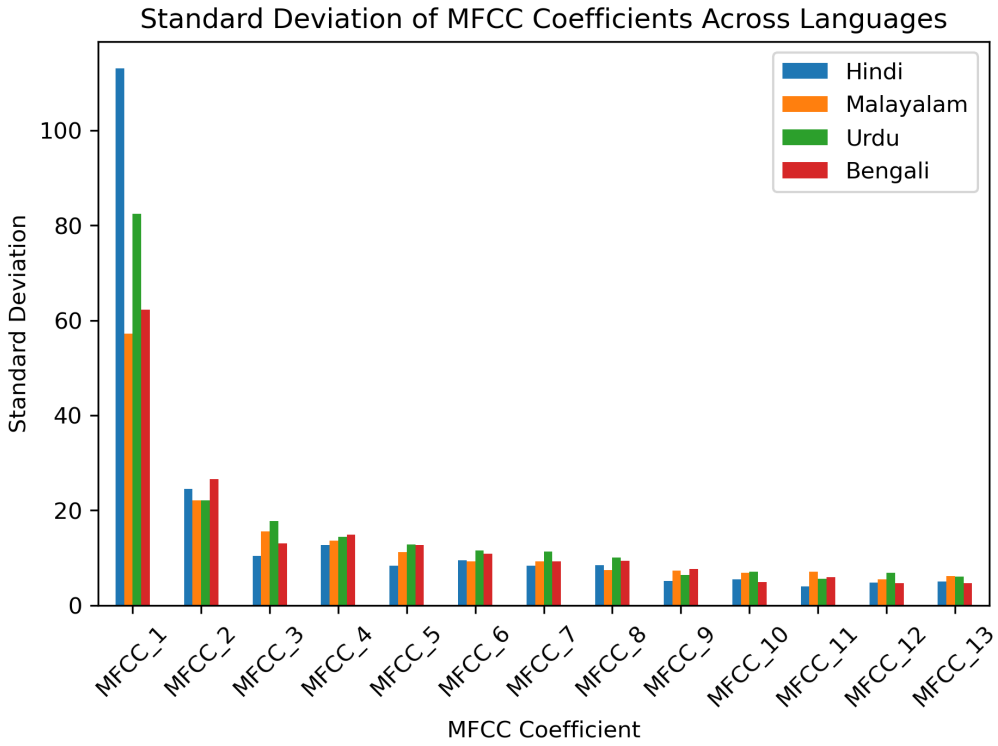

Figure 4: Standard deviation of MFCC coefficients across languages

Key observations from the statistical analysis include:

- MFCC_1 (representing overall energy) showed substantial negative values for all languages, with Urdu showing the least negative value (-316.04) and Malayalam the most negative (-367.47)
- MFCC_2 exhibited the highest positive values across all languages, with Urdu showing the maximum (117.35) and Hindi the minimum (76.72)
- Higher-order coefficients (MFCC_6 through MFCC_13) showed more subtle differences between languages
- Standard deviations were highest for MFCC_1 across all languages, with Hindi showing the greatest variability

These patterns indicated clear statistical differences in acoustic characteristics between languages, suggesting that the MFCC coefficients effectively capture language-specific characteristics.

3.3 Statistical Significance Testing

One-way ANOVA was performed for each MFCC coefficient to test the null hypothesis that all languages have the same mean value. The results are summarized in Table 1.

Coefficient	F-value	p-value	Significant
MFCC_1	317.1585	0.0000	Yes
MFCC_2	2293.3570	0.0000	Yes
MFCC_3	668.5543	0.0000	Yes
MFCC_4	1442.7287	0.0000	Yes
MFCC_5	511.5516	0.0000	Yes
MFCC_6	40.1060	0.0000	Yes
MFCC_7	5347.3550	0.0000	Yes
MFCC_8	187.2923	0.0000	Yes
MFCC_9	1240.1822	0.0000	Yes
MFCC_10	1906.1025	0.0000	Yes
MFCC_11	430.2546	0.0000	Yes
MFCC_12	2392.0471	0.0000	Yes
MFCC_13	373.0259	0.0000	Yes

Table 1: ANOVA results for MFCC coefficients across languages

The ANOVA results showed that all 13 MFCC coefficients exhibited statistically significant differences across languages ($p < 0.0001$), with particularly high F-values for MFCC_7 ($F = 5347.36$), MFCC_12 ($F = 2392.05$), and MFCC_2 ($F = 2293.36$). This confirmed that the languages have distinct acoustic profiles as captured by the MFCC features.

To further investigate which language pairs showed significant differences, pairwise t-tests were conducted. The results are visualized in Figure 5.

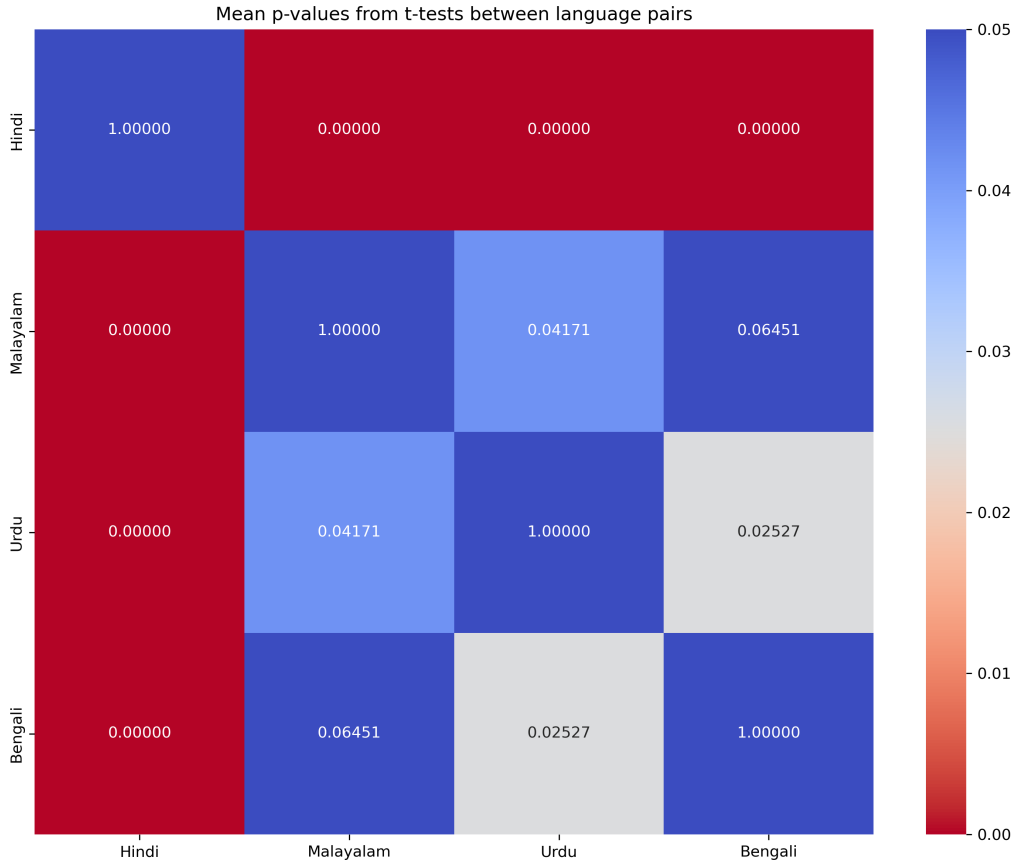

Figure 5: Mean p-values from t-tests between language pairs

The pairwise t-test results revealed:

- Hindi was significantly different from all other languages ($p < 0.0001$)
- Malayalam and Urdu showed significant differences ($p = 0.041711$)
- Malayalam and Bengali did not show statistically significant differences ($p = 0.064510$)
- Urdu and Bengali exhibited significant differences ($p = 0.025274$)

This analysis indicated that while most language pairs could be distinguished based on their MFCC features, Malayalam and Bengali showed more acoustic similarities. This finding suggested potential challenges in correctly differentiating between these two languages in a classification task.

4 Language Classification Models

Building upon the MFCC analysis, machine learning models were implemented to classify audio samples by language based on their MFCC features.

4.1 Model Development Methodology

4.1.1 Feature Preparation

The MFCC features were processed to create a suitable input format for classification models:

- Mean values of each MFCC coefficient across time were calculated, resulting in a 13-dimensional feature vector per sample
- Data was split into training (75%) and testing (25%) sets using stratified sampling to maintain class balance
- Features were standardized using scikit-learn's StandardScaler [6]

This process transformed the original time-frequency MFCC representation into a compact feature set while preserving the distinctive characteristics of each language.

4.1.2 Classification Models

Three different classification models were implemented and evaluated:

- **Support Vector Machine (SVM)** with RBF kernel
- **Random Forest** with 100 estimators
- **Neural Network** (Multi-layer Perceptron) with a single hidden layer of 100 units

The models were selected to represent different classification approaches, from margin-based (SVM) to ensemble methods (Random Forest) and neural approaches (MLP).

4.2 Model Evaluation Results

The models were trained on 12,000 samples (3,000 per language) and evaluated on a held-out test set of 4,000 samples (1,000 per language). Figure 6 shows the comparative accuracy of the three models.

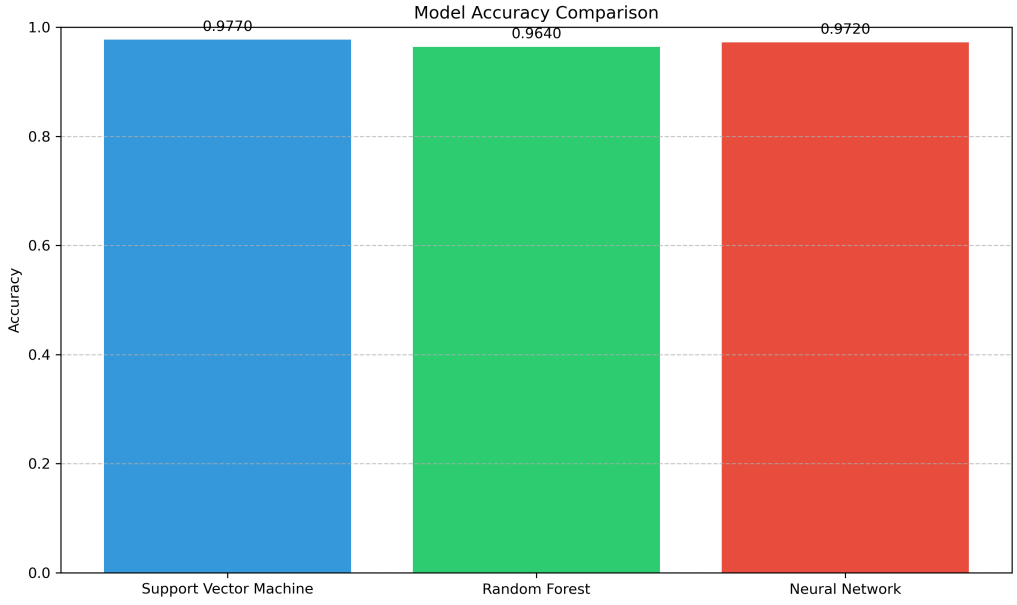

Figure 6: Comparison of model accuracies for language classification

The Support Vector Machine achieved the highest accuracy at 97.70%, closely followed by the Neural Network at 97.20%. The Random Forest classifier performed slightly worse with an accuracy of 96.40%.

4.2.1 Detailed Classification Performance

The classification report for the best-performing model (SVM) showed:

- **Bengali**: Precision = 0.96, Recall = 0.97, F1-score = 0.97
- **Hindi**: Precision = 0.99, Recall = 0.98, F1-score = 0.99
- **Malayalam**: Precision = 0.99, Recall = 0.98, F1-score = 0.98
- **Urdu**: Precision = 0.97, Recall = 0.98, F1-score = 0.97

These metrics indicate that the model performed exceptionally well across all four languages, with Hindi and Malayalam showing the highest precision.

4.2.2 Confusion Matrix Analysis

Figure 7 shows the confusion matrices for all three models, providing insight into the specific patterns of misclassification.

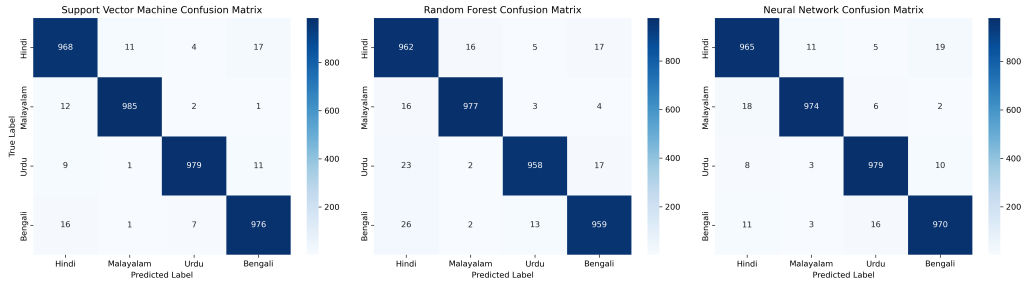

Figure 7: Confusion matrices for SVM, Random Forest, and Neural Network models

Key observations from the confusion matrices:

- All models showed some confusion between Bengali and Malayalam, which aligns with the findings from the pairwise t-tests showing these languages had the least significant acoustic differences
- Malayalam was the most accurately classified language across all models
- The SVM model showed fewer misclassifications overall compared to the other models

5 Discussion

5.1 MFCC as Language Discriminators

The statistical analysis of MFCC features demonstrated that these coefficients effectively capture the acoustic characteristics that distinguish different Indian languages. The high F-values in the ANOVA tests and the strong classification performance validate the use of MFCCs as a foundation for language identification systems.

The particular importance of lower-order MFCC coefficients (especially MFCC_2 and MFCC_7) suggests that fundamental acoustic properties like spectral envelope and energy distribution play a crucial role in differentiating languages. This aligns with linguistic theory, as these features often correspond to differences in phonetic inventory and articulatory dynamics across languages [3].

5.2 Classification Model Performance

The high accuracy achieved by all three models (96%) demonstrates that MFCC-based features provide sufficient information for reliable language identification. The superior performance of the SVM model suggests that the language classes, while not linearly separable, can be effectively distinguished in the higher-dimensional space created by the RBF kernel.

The Neural Network's comparable performance indicates that more complex architectures might not provide significant advantages for this specific task and feature set. This finding is valuable from a computational efficiency perspective, as SVMs generally require fewer resources for deployment compared to neural networks.

5.3 Challenges and Limitations

Despite the strong results, several challenges and limitations should be acknowledged:

- **Speaker Variability**: The dataset likely contains samples from multiple speakers, but speaker identity was not controlled for in this analysis. Individual speaking styles, accents, and vocal characteristics could influence the MFCC patterns independently of language.
- **Feature Representation**: Using mean MFCC values discards temporal dynamics that might contain important language-specific information. While this simplification proved effective, more sophisticated representations might improve performance further.
- **Language Similarity**: The analysis revealed acoustic similarities between Malayalam and Bengali, which could pose challenges if more closely related languages were included in the classification task.
- **Environmental Factors**: The classification performance might decrease in real-world applications where audio samples contain varying levels of environmental noise and recording conditions.

5.4 Future Work

Several directions for future work could address these limitations and extend the current findings:

- **Temporal Modeling**: Incorporating sequence models (e.g., RNNs, LSTMs) to capture the temporal dynamics of speech that are lost when using averaged features [2].
- **Feature Engineering**: Exploring additional acoustic features beyond MFCCs, such as prosodic features (pitch, intonation, rhythm) or spectral features (centroid, flux, roll-off) [1].
- **Robustness Testing**: Evaluating model performance under various noise conditions and with speech from different regional accents to assess real-world applicability.
- **Expanded Language Set**: Including more Indian languages, particularly those with known linguistic similarities, to test the limits of MFCC-based discrimination.
- **Cross-Speaker Validation**: Implementing speaker-independent evaluation to ensure that models are learning language characteristics rather than speaker-specific patterns.

6 Conclusion

This task has successfully demonstrated the effectiveness of MFCC features for distinguishing between four Indian languages: Hindi, Malayalam, Urdu, and Bengali. The statistical analysis revealed significant acoustic differences between most language pairs, providing a solid foundation for automated language identification.

The classification models, particularly the Support Vector Machine, achieved high accuracy rates (up to 97.7%), confirming that MFCC-based features enable reliable language identification. This performance is particularly impressive considering the simplified feature representation used (mean values of MFCC coefficients), suggesting that even more sophisticated approaches could further improve results.

The findings have important implications for multilingual speech processing applications in India and other linguistically diverse regions. Automated language identification systems based on MFCC features could enhance accessibility of speech-based services, improve machine translation workflows, and support language documentation efforts.

Future work should focus on addressing the identified limitations, particularly by incorporating temporal modeling, expanding the language set, and testing robustness under varied conditions. These extensions would further strengthen the foundation established by this study and move closer to practical deployment of MFCC-based language identification systems.

References

[1] Giannakopoulos, T. and Pikrakis, A. (2009). Audio features for speech/music discrimination in radio broadcasts. In *Proceedings of the International Conference on Digital Audio Effects*.

[2] Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE.

[3] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR.

[4] Kaggle (2023). Language detection dataset. https://www.kaggle.com/datasets/toponowicz/spoken-language-identification.

[5] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, volume 8, pages 18–25.

[6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

[7] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386.

[8] Ittichaichareon, C., Suksri, S., and Yingthawornsuk, T. (2012). Speech recognition using MFCC. In *International Conference on Computer Graphics, Simulation and Modeling (ICGSM)*, pages 28–29, Pattaya, Thailand.