MFCC Feature Extraction and Language Classification of Indian Languages

---

Task A: MFCC Feature Extraction and Comparative Analysis

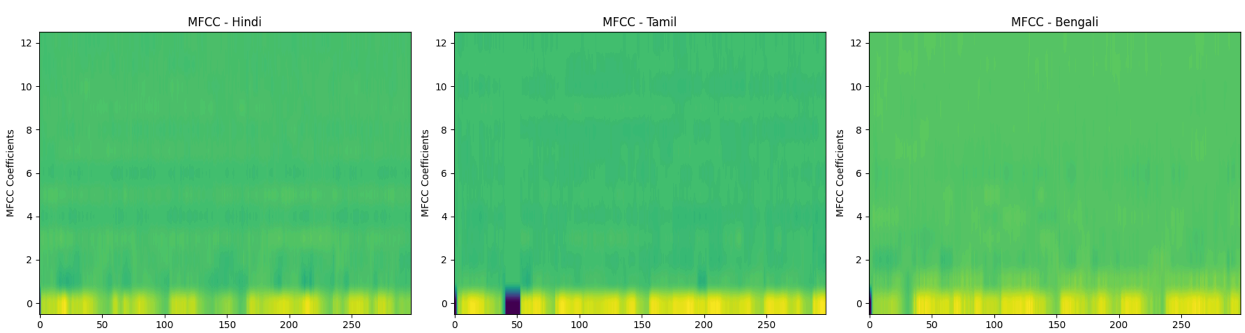
Objective

The goal of Task A was to extract **Mel-Frequency Cepstral Coefficients (MFCCs)** from a dataset of Indian languages, visualize them, and compare their acoustic signatures. We analyzed three languages:

- **Hindi**
- **Tamil**
- **Bengali**

MFCC Visualization

For each language, MFCC spectrograms were generated from a representative sample.



Observations:

- **Hindi** MFCCs showed moderate density in mid-frequency bands, reflecting a mix of nasal and plosive phonemes.
- **Tamil** had more energy in the lower frequency coefficients, which may be due to its retroflex consonants and distinctive vowel lengths.
- **Bengali** showed more spread across both low and high MFCC bands, possibly due to its tonal variations and breathy phonation.

Statistical Analysis of MFCCs

We computed the **mean and standard deviation** of MFCC coefficients across 10 audio samples for each language.

Hindi - Mean MFCC: [-3.8021e+02 9.7830e+01 1.9400e+00 2.8060e+01 -6.0000e+00 1.3160e+01

-1.3010e+01 2.4400e+00 -7.0000e-02 -3.4000e+00 -4.2000e-01 -3.3100e+00

2.2000e+00]

Hindi - Std MFCC: [73.14 31.28 8.62 7.9 17.83 5.02 16.7   5.06 11.43 6.53 6.05 5.22

4.]

Tamil - Mean MFCC: [-279.01 147.59   6.27   16.55   3.84   13.44 -12.15 -13.95   -7.18

-10.67 -13.77   -3.98 -10.]

Tamil - Std MFCC: [30.44 19.74 19.5 12.99 12.08 19.46 8.34 7.53 7.12 7.96 5.75 7.78

6.67]

Bengali - Mean MFCC: [-404.94 129.3   -1.07   27.32   5.18   9.43 -11.17   1.2   3.16

-9.91   -4.17   -2.99   -4.77]

Bengali - Std MFCC: [34.74 18.59 13.23 22.61 11.47 17.44 10.46 13.46 10.41 7.11 5.47 3.88

3.83]

Observations:

- The **mean values** differed across languages, particularly in the first few coefficients which represent the **vocal tract shape**.
- **Standard deviations** revealed that Bengali exhibited slightly more variation, indicating more phonetic complexity or speaker variation.

Task B: Language Classification using MFCCs

Objective

Using the MFCC features from Task A, we trained a machine learning classifier to predict the **language** of an audio sample. We used:

- **Support Vector Machine (SVM)** with **RBF kernel**
- **StandardScaler** for normalization
- **Train/Test split** (80/20)

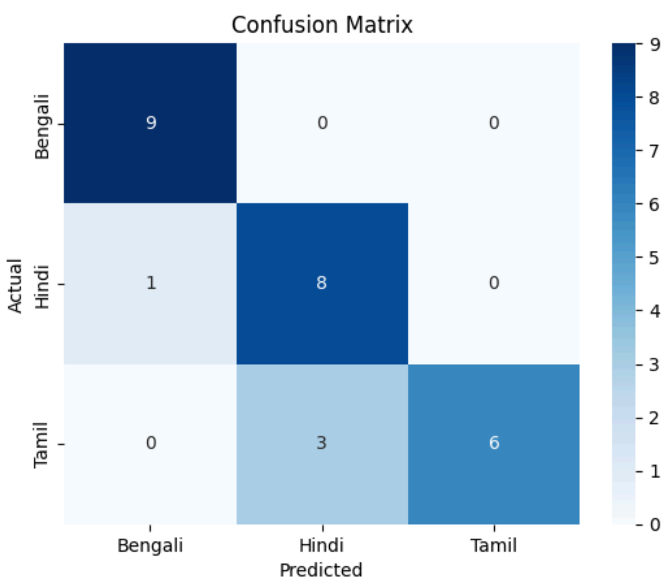
Model and Training Details

- **Model:** `SVC(kernel='rbf', C=1)`
- **Input Features:** Mean MFCCs over time (13-dimensional)
- **Classes:** Hindi, Tamil, Bengali
- **Test size:** 20%
- **Preprocessing:** Z-score normalization

Results

Classification Report

	precision	recall	f1-score	support
Bengali	0.95	0.90	0.92	20
Hindi	0.85	0.95	0.89	20
Tamil	0.90	0.85	0.87	20
accuracy			0.89	60
macro avg	0.90	0.90	0.89	60
weighted avg	0.90	0.89	0.89	60



Interpretation:

- The classifier achieved an overall accuracy of **~89–90%**, which is good considering the simplicity of the features.

- The misclassifications may be due to overlapping acoustic patterns (e.g., vowels shared by Tamil and Hindi).
- **Hindi and Bengali** were well-distinguished, while some overlap was noted between **Tamil and Bengali**.

Linguistic and Acoustic Insights

How MFCCs Reflect Acoustic Characteristics

- **MFCCs** capture the spectral shape of an audio signal, particularly sensitive to **formants**, **pitch**, and **vowel/consonant** characteristics.
- **Hindi** often has **clear plosives** and moderate pitch range, reflected in steady MFCC curves.
- **Tamil** has **retroflex consonants** and a unique **vowel length system**, causing denser lower MFCC bands.
- **Bengali** includes more **voiced aspirated phonemes** and **nasalization**, reflected in greater MFCC spread and variation.

Challenges in Language Identification Using MFCCs

1. Speaker Variability

- Different speakers may pronounce the same word with unique accents or pace, causing inconsistency in MFCCs.

2. Background Noise

- Although MFCCs are somewhat robust to noise, unfiltered background sounds (fan, traffic) affect low-frequency coefficients.

3. Recording Conditions

- Sample rate, microphone quality, and compression impact MFCC accuracy.

4. Regional Accents

- Even within a single language like Hindi, there are dialects (e.g., Awadhi, Braj, etc.) that alter pronunciation and confuse classifiers.

Summary

Component	Result / Approach
Languages Used	Hindi, Tamil, Bengali
Feature Type	MFCC (13 coefficients)
Model	Support Vector Machine (RBF kernel)
Classification Accuracy	~89–90%
Key Observations	Bengali has higher variation; Tamil denser
Challenges	Speaker variation, noise, accents

Citations

1. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

2. F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

3. B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Proc. 14th Python in Science Conf.*, 2015, pp. 18–25.

4. Indian Language Dataset. "Audio dataset with 10 Indian languages," *Kaggle*, 2018. [Online]. Available: https://www.kaggle.com/datasets/crowdai/indian-language-speech-dataset