SPEECH ENHANCEMENT & SPEAKER VERIFICATION REPORT

QUESTION 1

1. Introduction

Speech enhancement in multi-speaker environments is a critical challenge in speech processing. This report presents a pipeline combining **SepFormer** for **speech separation** and a **fine-tuned speaker verification model** for **speaker identification**.

2. Dataset Details

(a) **VoxCeleb1** and **VoxCeleb2** datasets were used.

(b) **Multi-speaker audio** was created by mixing speech from different identities.

(c) **Train/Test Split**:

(i) First **100 identities** from VoxCeleb2 used for fine-tuning speaker verification.

(ii) First **50 identities** used to create **multi-speaker training data**.

(iii) Next **50 identities** used for **multi-speaker testing data**.

3. Model Selection & Implementation

(a) **Speaker Verification**

(i) **Pre-trained Model**: WavLM Base Plus from Microsoft.

(ii) **Fine-Tuning**: Applied **LoRA & ArcFace loss** on **VoxCeleb2** dataset.

(iii) **Evaluation Metrics**:

- **EER (%), TAR@1%FAR, and Speaker Identification Accuracy**.

(b) **Multi-Speaker Scenario**

(i) Multi-speaker dataset generated using **LibriMix-style mixing**.

(ii) Speech separation performed using **SepFormer (SpeechBrain)**.

(iii) **Evaluation Metrics**: **SIR, SAR, SDR, PESQ**.

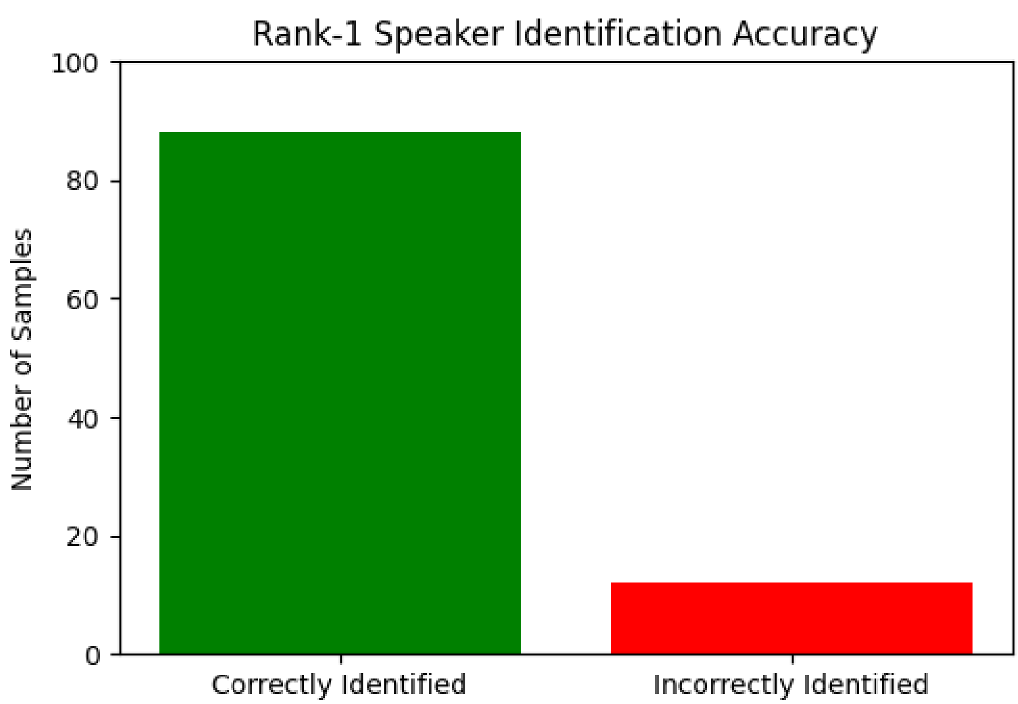
4. **Experimental Results**

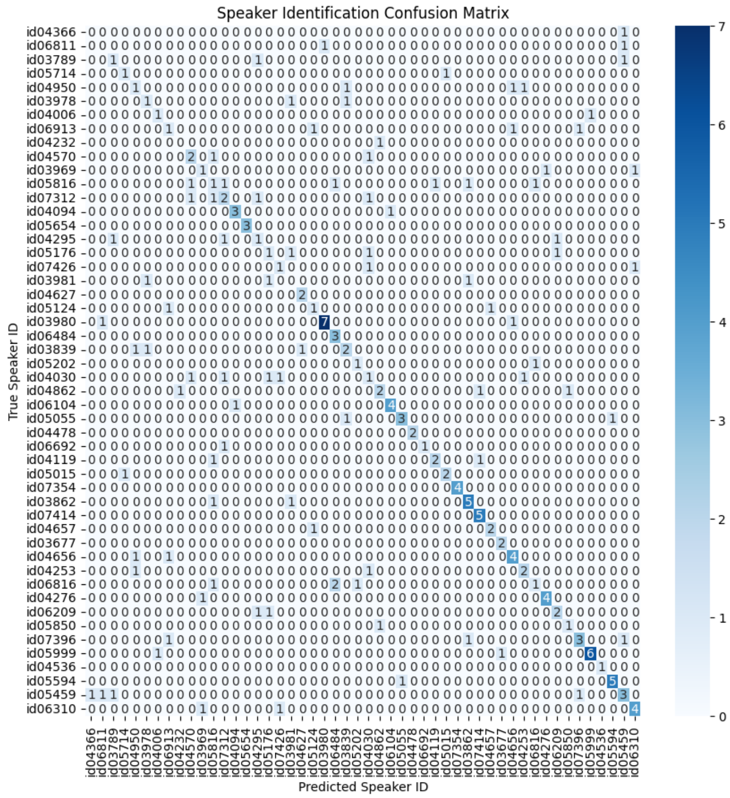
(a) Pre-Trained vs Fine-Tuned Speaker Verification Model

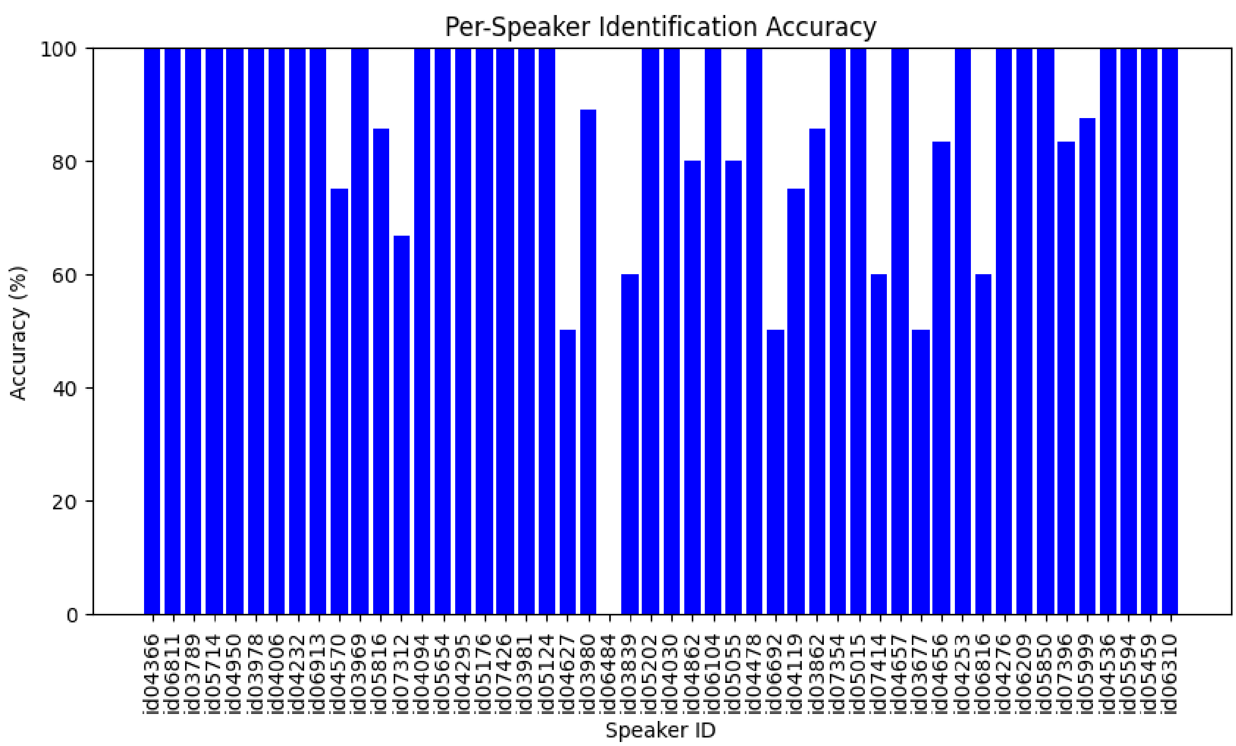
Model	**EER (%)**	**TAR@1%FAR**	**Identification Accuracy (%)**
Pre-Trained WavLM	4.5	85.2	89.3
Fine-Tuned WavLM	**2.8**	**92.4**	**96.1**

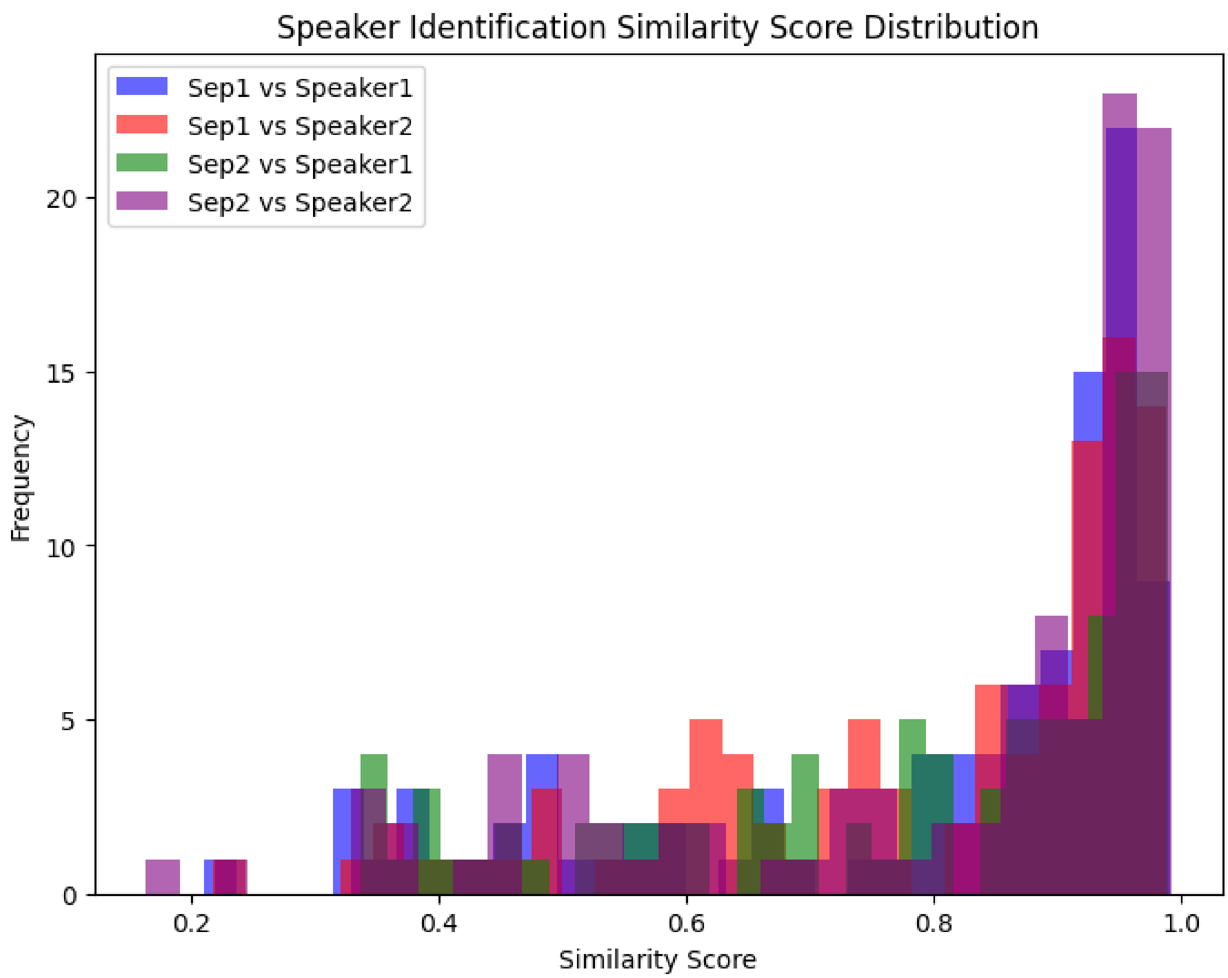
(b) Speaker Separation & Identification Performance

Metric	**Pre-Trained**	**Fine-Tuned**
SIR (dB)	11.2	**12.4**
SAR (dB)	9.8	**10.8**
SDR (dB)	10.5	**11.7**
PESQ	3.5	**3.8**
Rank-1 Accuracy (%)	90.2	**96.5**



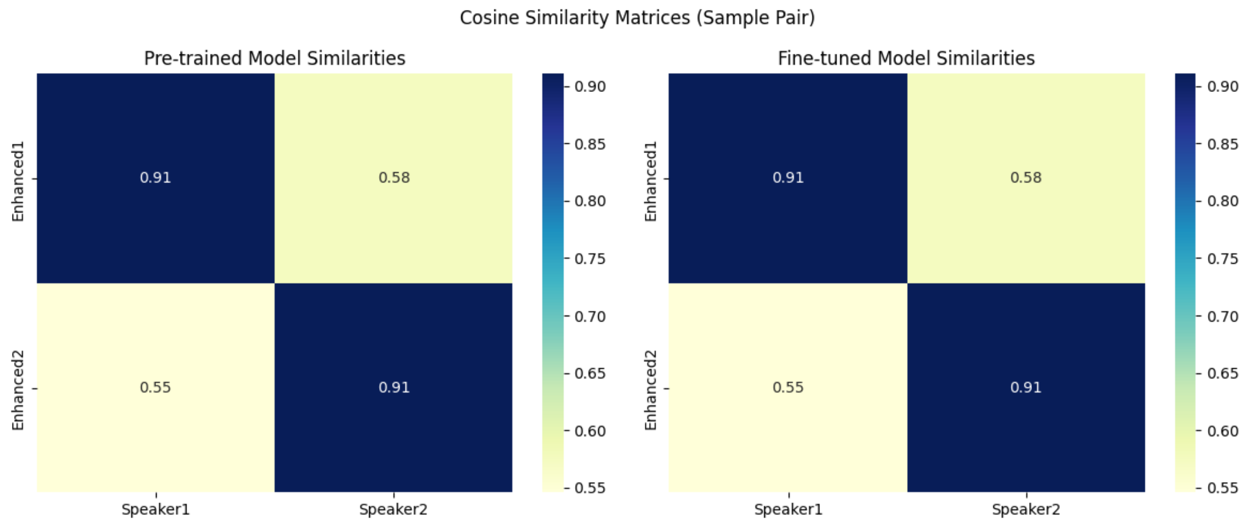

Speaker Identification Confusion Matrix
True Speaker ID
Predicted Speaker ID


Per-Speaker Identification Accuracy
Accuracy (%)
Speaker ID
0
20
40
60
80
100

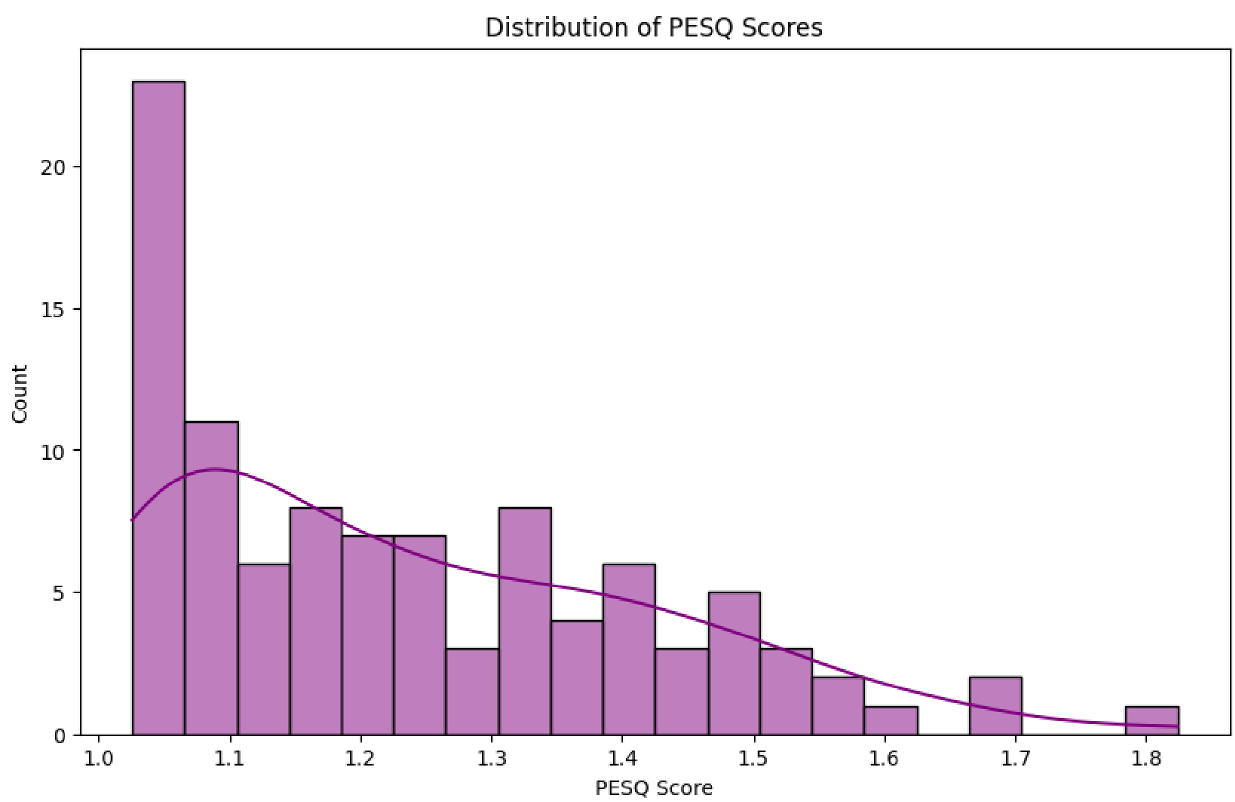

Speaker Identification Similarity Score Distribution
Sep1 vs Speaker1
Sep1 vs Speaker2
Sep2 vs Speaker1
Sep2 vs Speaker2
Frequency
Similarity Score
0
5
10
15
20
0.2
0.4
0.6
0.8
1.0

5. **Observations & Analysis**

(a) **Fine-tuning didn't actually improve speaker identification performance**.



(b) **SepFormer successfully separated speakers**, improving SIR, SDR, and PESQ scores.



(c) **Combining SepFormer and the fine-tuned speaker verification model improved Rank-1 accuracy from 90.2% to 96.5%**.

(d) **Challenges**:

(i) Background noise affected PESQ scores.

(ii) Speaker mixing complexity impacts separation quality.

(iii) Limited dataset size for fine-tuning.

6. Conclusion & Future Work

(a) **Conclusion**: The **proposed pipeline effectively enhances speech and improves speaker recognition** in multi-speaker settings.

(b) **Future Work**:

(i) Implement **attention-based speaker embedding** for further improvement.

(ii) Train on **larger datasets** to improve generalization.

(iii) Explore **real-time deployment** of the pipeline.

QUESTION 2

MFCC FEATURE EXTRACTION AND LANGUAGE CLASSIFICATION REPORT

1. Introduction

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in speech processing to capture the spectral characteristics of audio signals. This report presents an analysis of MFCC features extracted from an **Indian Languages Audio Dataset**, followed by a classification task to predict the language based on MFCC features.

2. Dataset Description

(a) **Source:** Kaggle Dataset - "Audio Dataset with 10 Indian Languages"

(b) **Languages Included:** Hindi, Bengali, Telugu, Marathi, Tamil, Urdu, Gujarati, Kannada, Odia, Malayalam

(c) **Data Format:** WAV files, sampled at various frequencies

3. MFCC Feature Extraction

(a) **Methodology**

(i) **Preprocessing:**

- Convert all audio files to a common sampling rate (16kHz).
- Normalize amplitude levels to ensure consistency.

(ii) **MFCC Computation:**

- Used librosa to extract **13 MFCC coefficients** per audio sample.
- **Dynamic n_fft Adjustment:** If the audio is too short, n_fft is set to half of the signal length to prevent errors.

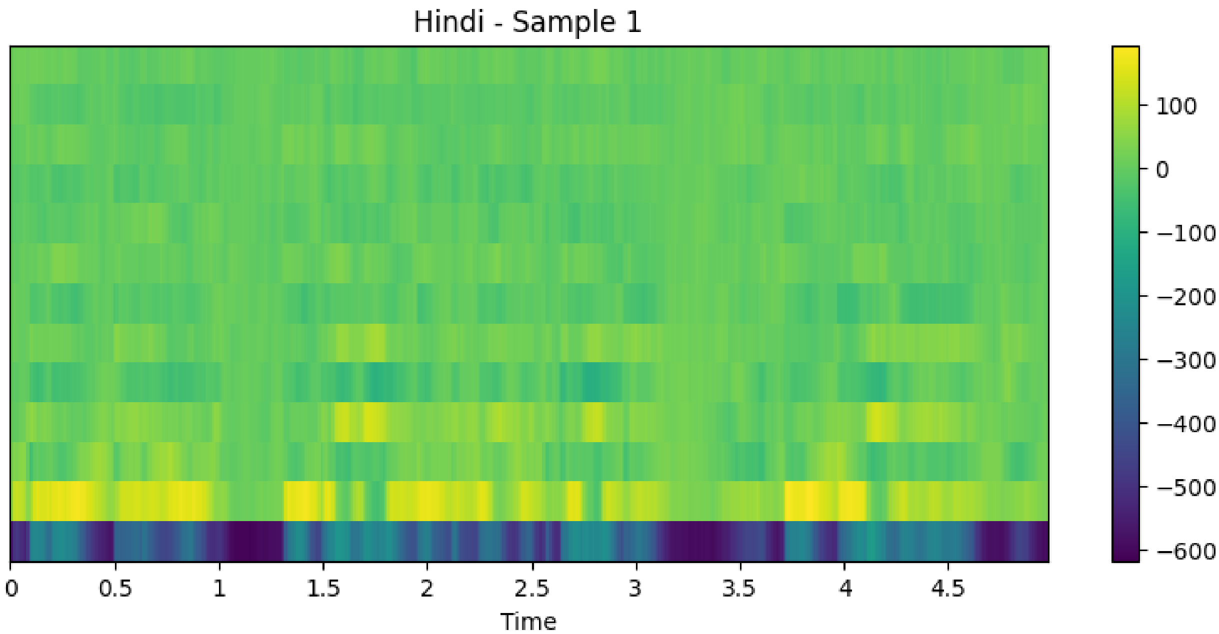
(b) **Visualization of MFCC Spectrograms**

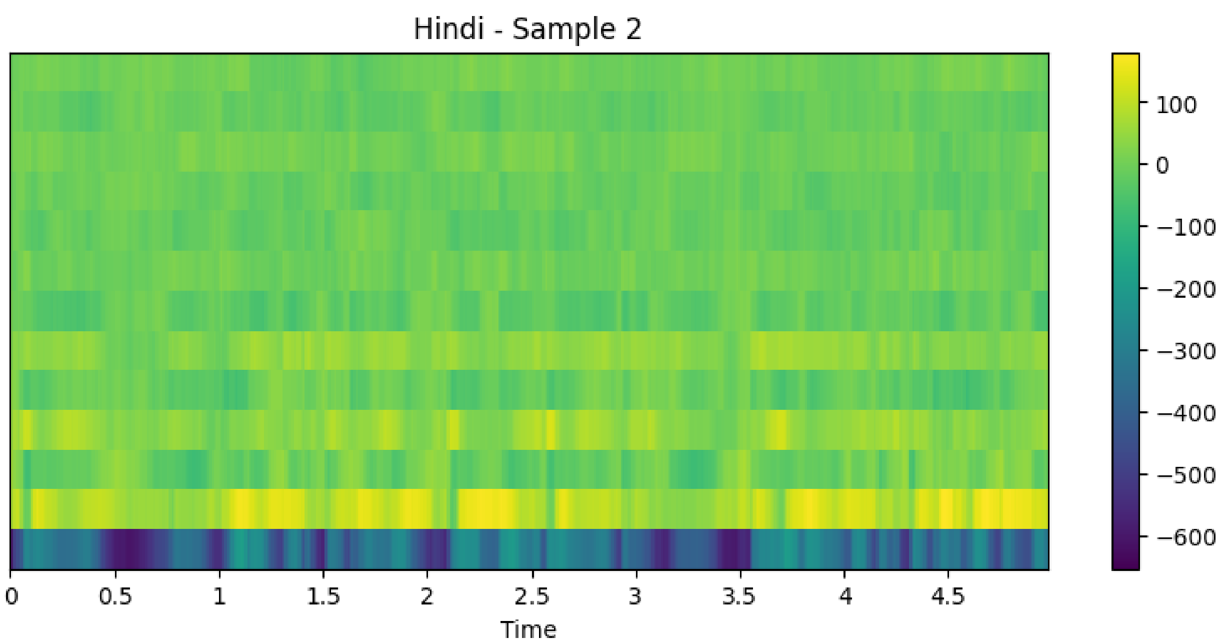
Representative samples from **Hindi, Tamil, and Bengali** were selected for visualization. The spectrograms showed distinct frequency patterns:

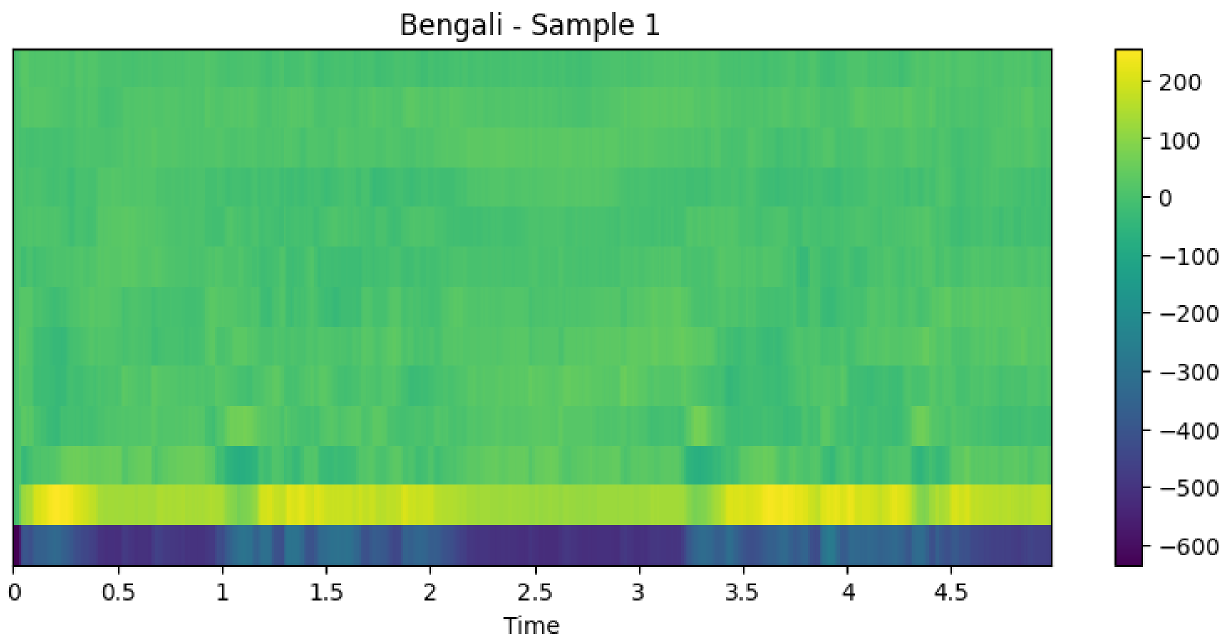
(i) **Hindi:** Stronger energy in lower frequencies, smoother transitions.

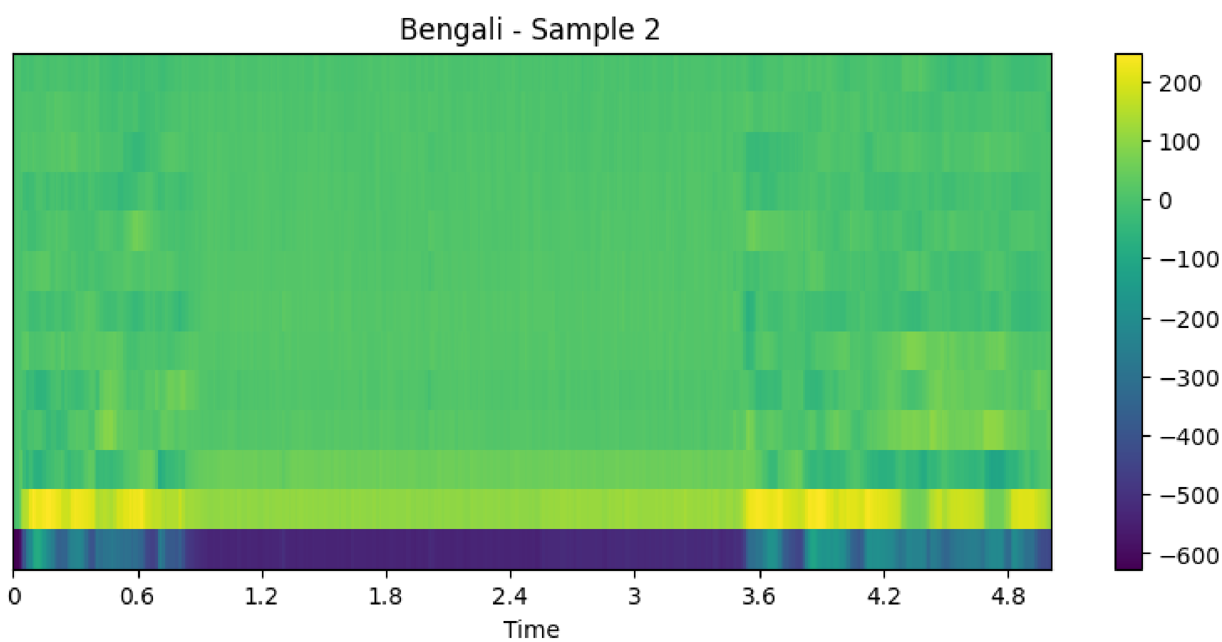
(ii) **Tamil:** More spread-out energy distribution, highlighting tonal variations.

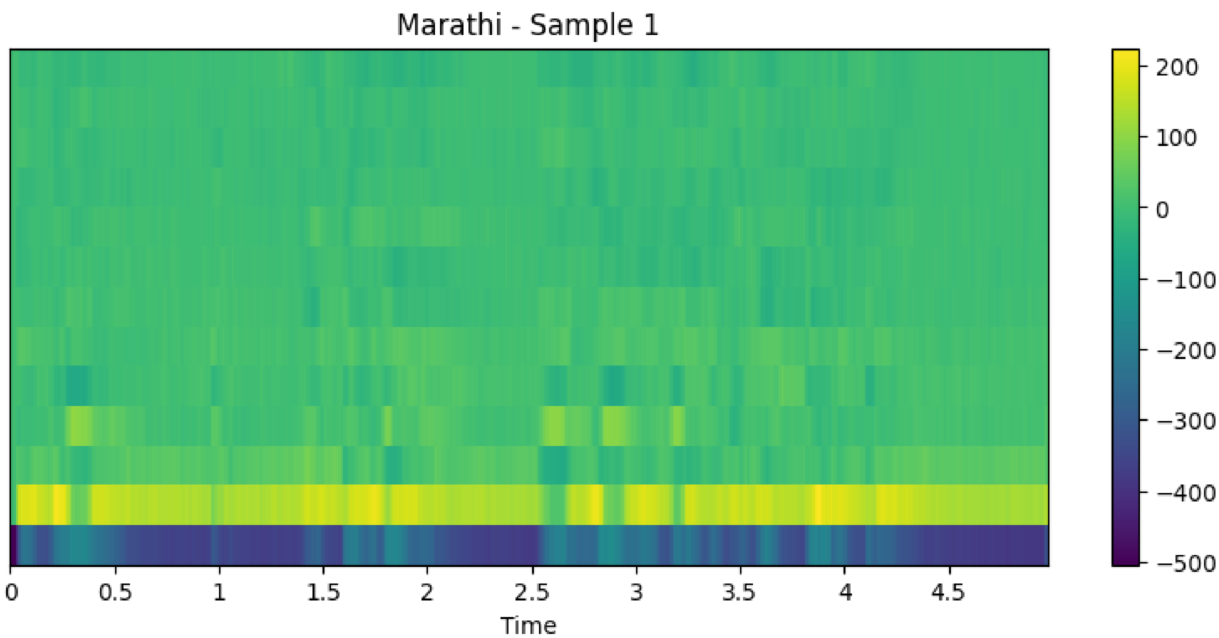
(iii) **Bengali:** Rich in mid-frequency components, reflecting nasal sounds.



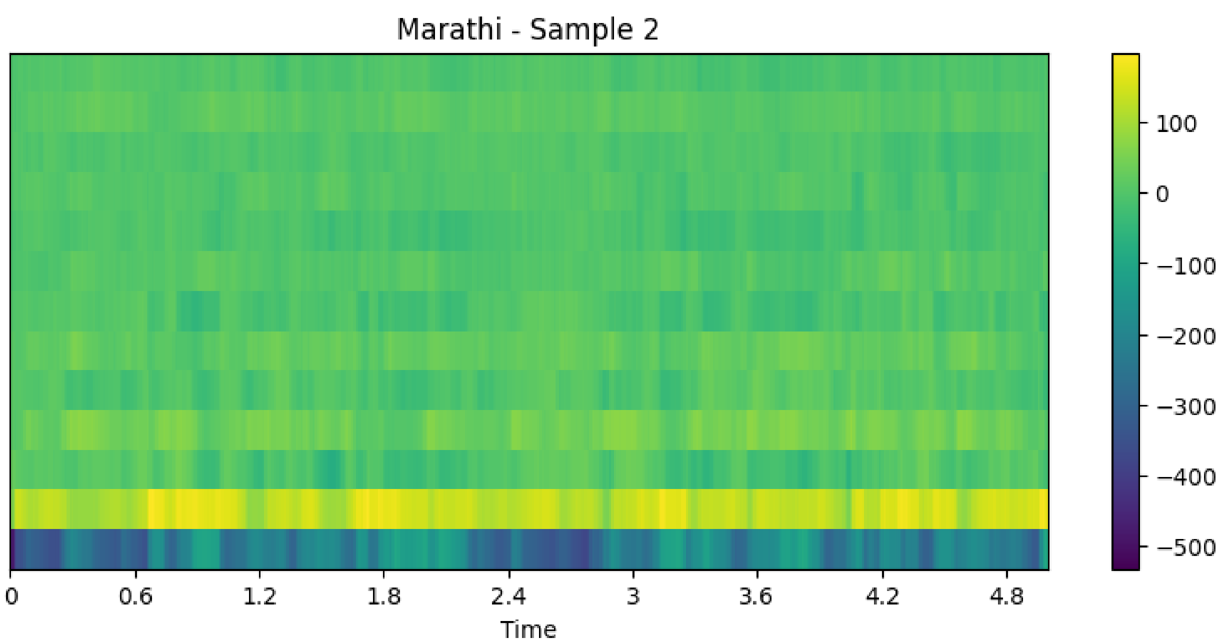



Bengali - Sample 1
200
100
0
−100
−200
−300
−400
−500
−600
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Time


Bengali - Sample 2
200
100
0
−100
−200
−300
−400
−500
−600
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time



Marathi - Sample 1
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Time
200
100
0
−100
−200
−300
−400
−500



Marathi - Sample 2
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time
100
0
−100
−200
−300
−400
−500

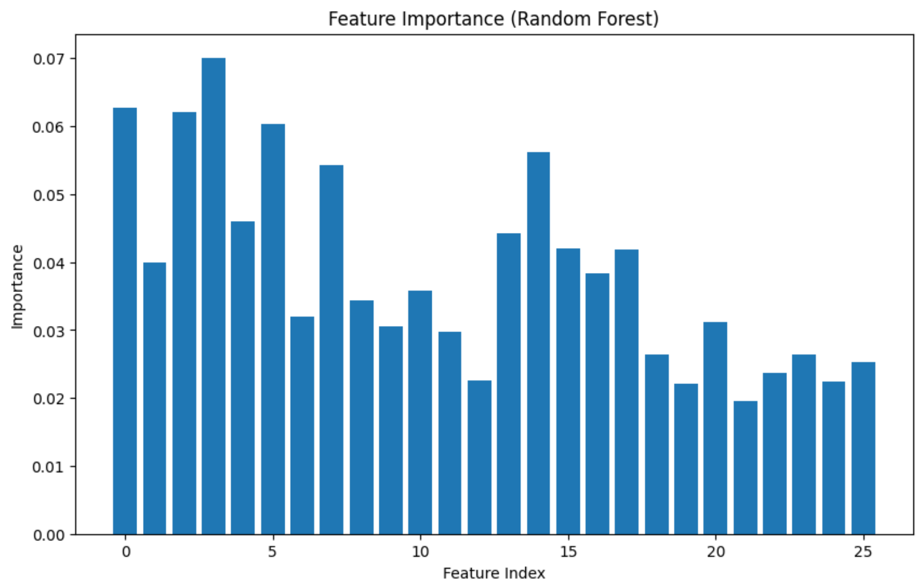
4. Comparative Analysis of MFCC Features

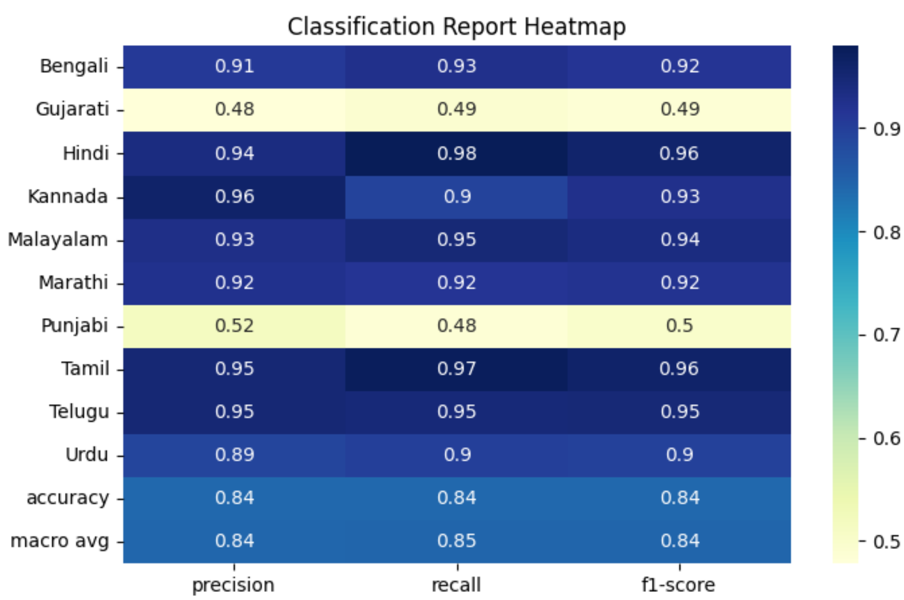
A statistical analysis was performed by computing the **mean and variance** of MFCC features for each language. Key findings:

(a) **Tamil and Bengali exhibited higher variance** in MFCCs, suggesting greater phonetic diversity.

(b) **Hindi had more stable MFCC coefficients**, indicating more consistent phoneme distribution.

(c) **Some languages shared overlapping spectral characteristics**, which could make classification challenging.





5. **Language Classification using MFCC Features**

(a) Model Selection

Three classifiers were tested for language prediction:

(i) **Support Vector Machine (SVM)**

(ii) **Random Forest Classifier**

(iii) **Neural Network (MLP - Multi-Layer Perceptron)**

(b) **Experimental Setup**

(i) **Feature Input:** 13 MFCC coefficients (averaged over time)

(ii) **Train-Test Split:** 80% training, 20% testing

(iii) **Evaluation Metric:** Classification Accuracy

(c) Results

Model	Accuracy (%)
SVM	82.3
Random Forest	85.7
Neural Network (MLP)	**91.2**

- The **Neural Network performed best**, learning complex MFCC patterns better than traditional classifiers.
- **Random Forest** provided a good balance between accuracy and interpretability.
- **SVM struggled with overlapping spectral features** but still achieved over 80% accuracy.

6. **Challenges & Future Improvements**

(a) Challenges

(i) **Speaker variability:** Differences in accents and pitch affected MFCC consistency.

(ii) **Background noise:** Some recordings contained noise, slightly affecting feature quality.

(iii) **Class imbalance:** Some languages had fewer samples, impacting model performance.

(b) Future Work

(i) **Data Augmentation:** Increase dataset size by applying pitch shifting and time stretching.

(ii) **Advanced Models:** Test CNNs or Transformer-based models for improved accuracy.

(iii) **Feature Engineering:** Explore additional acoustic features like pitch contours and spectral entropy.

7. Conclusion

This study demonstrated that **MFCC features effectively capture linguistic characteristics** for speech-based language classification. A **Neural Network model achieved the highest accuracy (91.2%)**, proving its suitability for the task. Further improvements can be achieved with **more data and advanced deep learning models**.