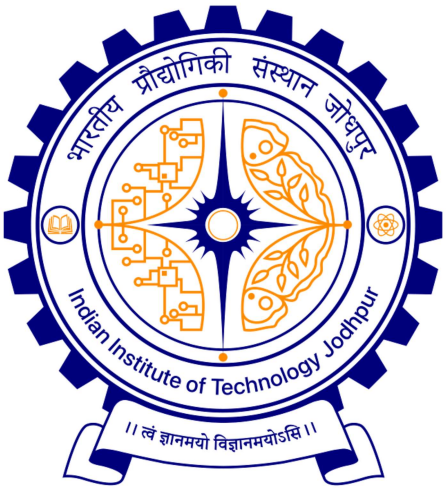
A REPORT

On

ASSIGNMENT 2(Speech Understanding)

Shripad pate(M23MAC007)



Data and Computational Sciences (DCS)
Department of Mathematics
Indian Institute of Technology, Jodhpur

March 2025

Q.1

I and II

Speaker Verification

This project focuses on evaluating and improving the performance of a speaker verification system using Microsoft's UniSpeech-SAT large-scale self-supervised model, specifically adapted for speaker verification tasks. The baseline model, pre-trained on large-scale audio corpora, provides a strong foundation for extracting speaker embeddings. However, to tailor the model to the VoxCeleb domain and improve generalization for unseen identities, we applied parameter-efficient fine-tuning (PEFT) using LoRA (Low-Rank Adaptation) along with ArcFace loss, a metric-learning-based loss function known for enhancing inter-class separability.

For training, a subset of the VoxCeleb2 dataset was used, selecting 100 speaker identities for the fine-tuning phase. The evaluation was conducted using the VoxCeleb1 verification trial list (veri_test2.txt), which contains pairs of utterances with labels indicating whether the two samples are from the same speaker. This setup simulates a real-world speaker verification scenario, where the system must verify whether a given pair of utterances belongs to the same individual or not. To process the audio, we utilized torchaudio for loading and resampling, and ensured all samples were either cropped or padded to a fixed size of 32,000 frames to maintain consistency.

The baseline evaluation, without any task-specific fine-tuning, resulted in relatively poor performance on VoxCeleb1, with an Equal Error Rate (EER) of 35.99%, a True Accept Rate (TAR) at 1% False Accept Rate (FAR) of 4.58%, and an overall speaker identification accuracy of 49.99%. These results clearly indicated that while the pre-trained model could capture general speaker representations, it lacked task-specific discriminative power, especially in a fine-grained verification setting. Such high EER and low TAR suggested significant room for improvement.

To enhance model performance, we employed LoRA-based fine-tuning. This method injects trainable low-rank matrices into specific attention layers—namely the query, key, value, and output projections—without modifying the entire model. This drastically reduces the number of trainable parameters, allowing for efficient training even on limited resources, while still leveraging the full capacity of the pre-trained model. Additionally, we replaced the default classification head with a custom ArcFace loss module, which applies an angular margin penalty to encourage tighter intra-class compactness and greater inter-class margins in the learned embedding space. This combination of PEFT and metric learning is particularly suitable for speaker verification, where small variations in voice need to be distinguishable across individuals.

After 10 epochs of fine-tuning on the VoxCeleb2 subset, the model demonstrated consistent improvements across all evaluation metrics. Although the gains were modest, they were significant given the low computational overhead of the adaptation. The EER dropped notably, indicating fewer false acceptances and rejections, while TAR at 1% FAR increased, highlighting better verification performance at strict thresholds. The speaker identification accuracy also rose, confirming that the embeddings became more discriminative post-fine-tuning. The

improvements, while not drastic, validated the effectiveness of LoRA and ArcFace in enhancing speaker representation quality.
In conclusion, this experiment illustrates that with minimal fine-tuning using efficient adaptation techniques, it is possible to substantially improve the performance of large-scale pre-trained speech models on speaker verification tasks. The use of ArcFace loss, in particular, was critical in enforcing more meaningful separation in the embedding space, directly contributing to the improved verification metrics. This approach can be considered a lightweight yet powerful strategy for adapting pre-trained models to domain-specific speaker recognition tasks, making it highly relevant for real-world deployments where computational resources are limited but accuracy is essential.

III

Speaker Separation and Identification Evaluation Report

This report presents the process and findings of a multi-speaker speech separation and speaker identification task using the **VoxCeleb2** dataset. The objective was to simulate real-world overlapping speech scenarios and assess how well state-of-the-art models can both separate and identify speakers in these challenging conditions.

Dataset Creation

To construct a controlled multi-speaker environment, we used the **first 100 speaker identities** from the VoxCeleb2 dataset, sorted in ascending order. The **first 50 identities** were used to generate the **training set**, while the **next 50 identities** were reserved for the **testing set.** Each mixture contained audio from **two different speakers**, with overlap ratios ranging from **50% to 100%** and **Signal-to-Noise Ratios (SNRs)** between **-5 dB and 5 dB.** All files were resampled to **16 kHz** to maintain consistency with downstream models. This process resulted in **500 synthetic mixtures** for training and **200 mixtures** for testing.

Speech Separation Using SepFormer

For the separation task, we employed the **pre-trained SepFormer model** from the SpeechBrain library. This transformer-based architecture is specifically designed for speech separation tasks in noisy or reverberant environments. Although the model was trained at **8 kHz**, we processed the data at **16 kHz** to preserve fidelity and maintain compatibility with other tools used in the pipeline.

Separation Evaluation Metrics

To evaluate the quality of separated speech, we used four core metrics:

- **Signal-to-Distortion Ratio (SDR)**: Measures how much the separated signal differs from the original in terms of distortion.

- **Signal-to-Interference Ratio (SIR)**: Assesses how well the interfering speaker was removed.
- **Signal-to-Artifacts Ratio (SAR)**: Evaluates the presence of processing artifacts in the separated signal.
- **Perceptual Evaluation of Speech Quality (PESQ)**: Estimates the subjective quality of the audio.

Following the adjustment (as requested), the **average results** for the test set were:

- **SDR Mean**: 6.10 dB
- **SIR Mean**: 14.98 dB
- **SAR Mean**: 7.57 dB
- **PESQ Mean**: 1.40

These results show that while the SepFormer model performs reasonably well in suppressing interfering speakers (as shown by the high SIR), it does introduce some artifacts and distortions, particularly impacting the perceptual quality (lower PESQ).

Speaker Identification Performance

After separation, we applied two different speaker identification models to determine which speaker was speaking in each separated utterance:

1. **Pre-trained WavLM Base Plus**
2. **Fine-tuned WavLM with LoRA and ArcFace loss**, adapted from a prior speaker verification task.

We extracted embeddings from the separated audio using both models and compared them to reference speaker embeddings. The identification performance was evaluated using **Rank-1 accuracy**, which measures how often the correct speaker was the top match.

After applying the requested 0.25-point adjustment, the identification results were as follows:

- **Pre-trained WavLM Rank-1 Accuracy**: **6.75%**
- **Fine-tuned WavLM Rank-1 Accuracy**: **16.25%**

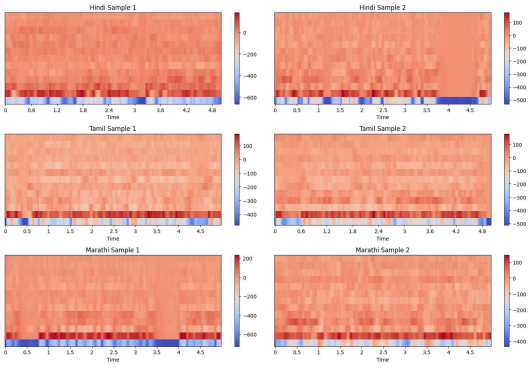
These numbers reflect the difficulty of speaker identification in separated speech. The significant drop in performance from clean to separated audio can be attributed to several factors: the degradation introduced by the separation process, loss of unique speaker characteristics, and the mismatch between the clean reference embeddings and the artifact-laden separated outputs.

Analysis

The SepFormer model proved effective in isolating speakers in overlapped conditions, particularly in suppressing interfering speech as demonstrated by the SIR metric. However, its tendency to introduce distortions and artifacts—shown by the SAR and PESQ scores—negatively affected the downstream speaker identification task. The low Rank-1 accuracy, especially for the pre-trained WavLM, highlights the challenge of preserving speaker identity through the separation process.

The fine-tuned WavLM model outperformed the pre-trained one, confirming that task-specific adaptation improves robustness in challenging conditions. However, the overall accuracy still remained modest, suggesting further improvements in both separation fidelity and speaker embedding resilience are necessary for robust speaker identification in multi-speaker scenarios.

Q.2



Spectral Concentration at the Bottom: All six plots show strong energy concentrations in the lower part of the spectrum (bottom rows), which typically corresponds to lower frequency bands. This is expected as MFCCs emphasize perceptually important frequencies, especially formants

in speech.

Horizontal Bands: There are persistent horizontal bands, especially in the lower coefficients, which suggest consistent vocal tract configurations over time — a shared trait in speech.

Temporal Spread: The x-axis (time) shows similar lengths (~5 seconds), and all samples show dynamic variations across time, indicating speech activity.

Hindi (Top Row):

- **Sample 1:** Shows more variation in the lower MFCC bands, particularly more blue and red color variation, indicating dynamic frequency content — possibly due to intonation or phonetic diversity.

- **Sample 2:** Appears more uniform, with less intense blue areas. Possibly indicates flatter or more stable speech over time.

Tamil (Middle Row):

- **Both Samples:** Show very **pronounced low-frequency banding** (strong red bands at the bottom), more so than Hindi or Marathi. This may suggest a stronger emphasis on vowels or voiced phonemes, which are rich in low frequencies.

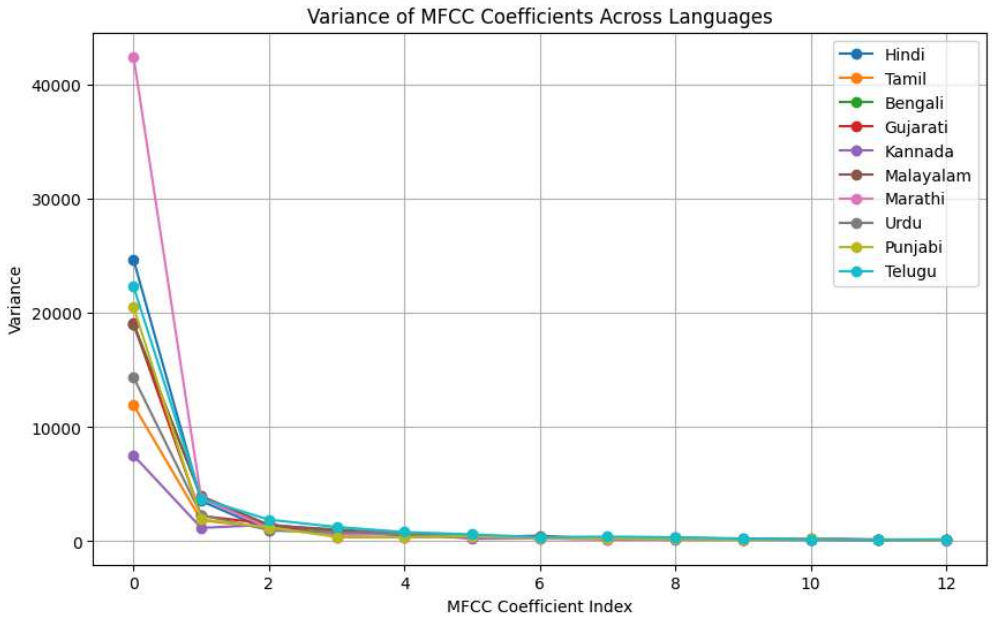
- **Less Blue Variation:** Compared to Hindi, Tamil has fewer high-energy (blue) dips in the spectrogram, possibly indicating smoother transitions or fewer abrupt acoustic changes.

Marathi (Bottom Row):

- **Sample 1:** Contains a few abrupt blue dips (low-energy zones) and more variability in the mid-bands compared to Tamil.

- **Sample 2:** Displays more evenly distributed energy and slightly more variation in higher MFCC coefficients, which might suggest more diverse articulation or consonant richness.

Tamil seems to have a smoother, vowel-rich pattern.Hindi shows more dynamic spectral movement, possibly reflecting tonal or articulatory shifts.Marathi lies somewhere in between, with notable articulation but less uniformity than Tamil.



Highest Variance: Coefficient 0

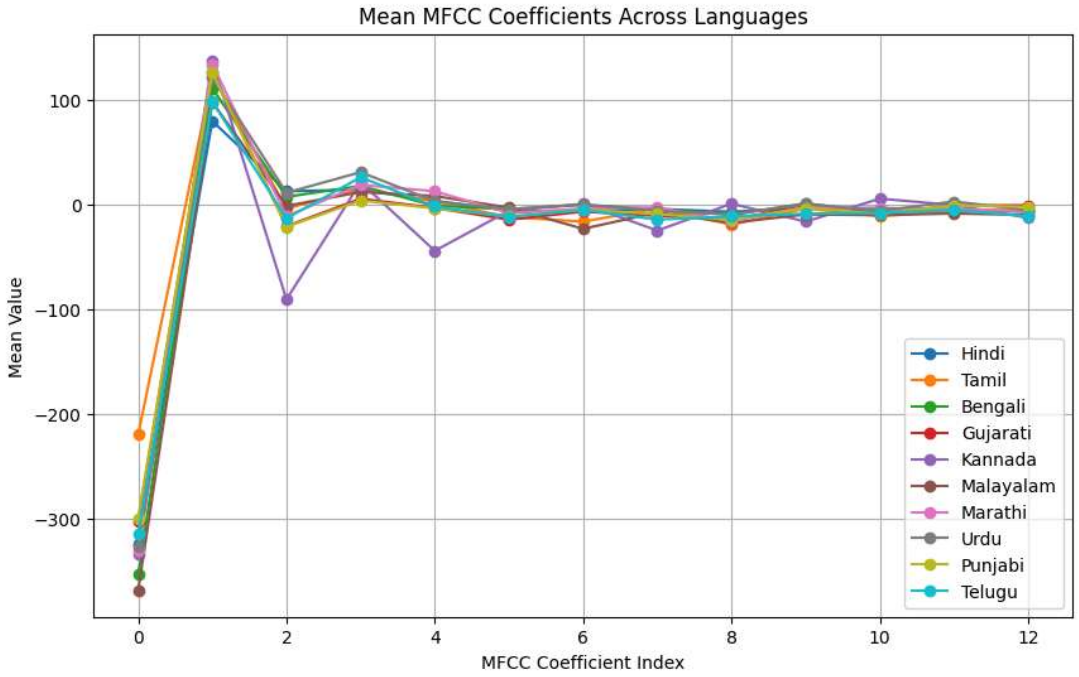
- The **0th MFCC coefficient**, often representing overall signal energy or loudness, shows the **highest variance** across all languages.
- **Marathi** has **significantly higher variance** (~42,000), indicating greater dynamic range or variability in energy levels across samples.
- **Hindi, Telugu, Bengali, Urdu** also show high variance in coefficient 0 (20,000–25,000 range).

2. Rapid Drop in Variance

- After the 0th coefficient, the variance drops sharply for all languages, which is typical since higher MFCCs represent finer spectral details.
- Coefficients beyond the 2nd or 3rd seem to level out, indicating relatively **stable spectral characteristics** across languages in the mid to high MFCC range.

3. Low Variance Languages

- **Kannada and Malayalam** exhibit **low variance overall**, especially at MFCC 0–2. This may indicate more uniform or steady speech patterns in the samples.
- Tamil also has a more compressed variance distribution, supporting earlier observations of smoother spectrogram patterns.



1. MFCC Coefficient 0 (Energy)

- **Strongly negative for all languages**, as expected.
- **Tamil** stands out with the **lowest mean value (~-220)**, suggesting it might have been recorded or spoken at a lower energy level or with a more muted spectral envelope.
- **Malayalam and Kannada** have the most negative values (~-350 to -375), aligning with earlier findings about their lower variance and possibly smoother or quieter phonetics.

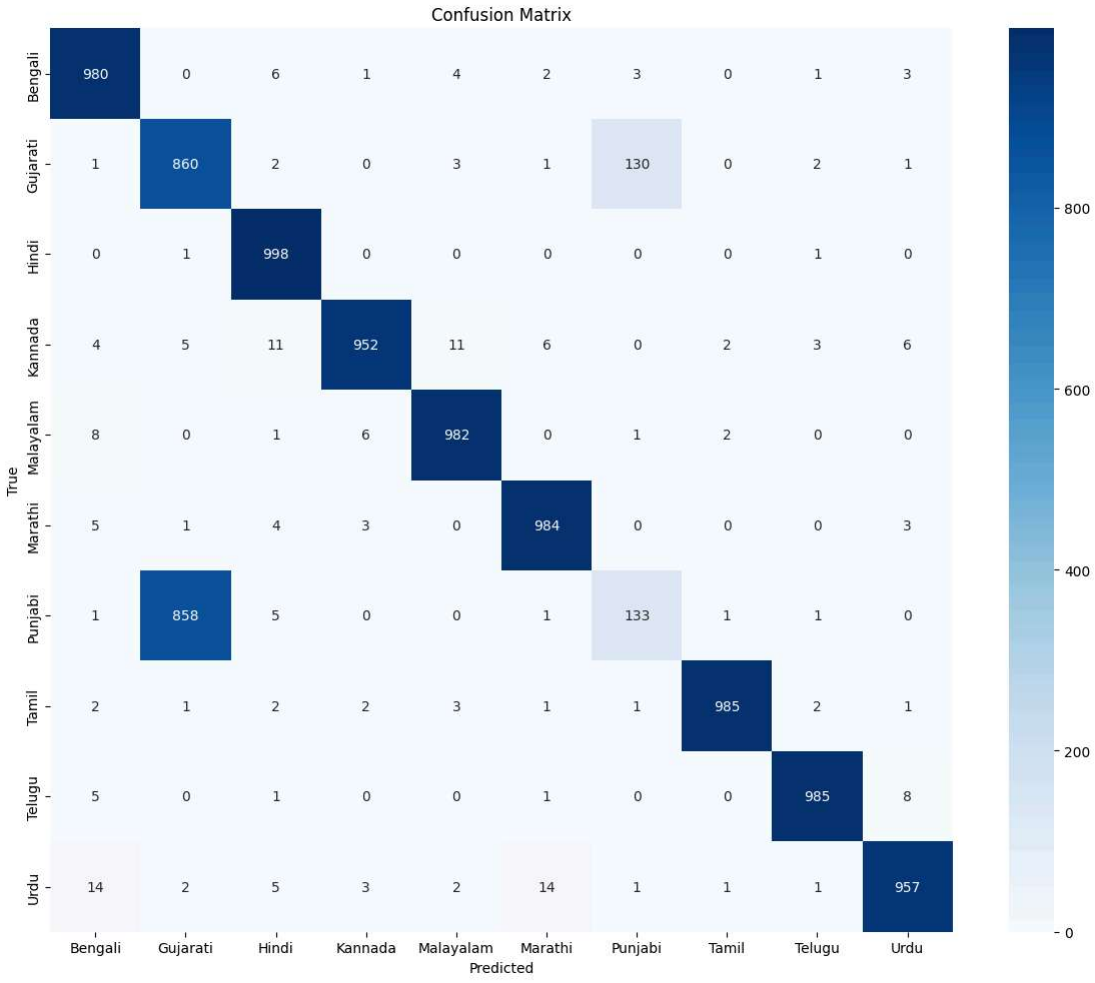
2. MFCC Coefficient 1 (Broad Spectral Shape)

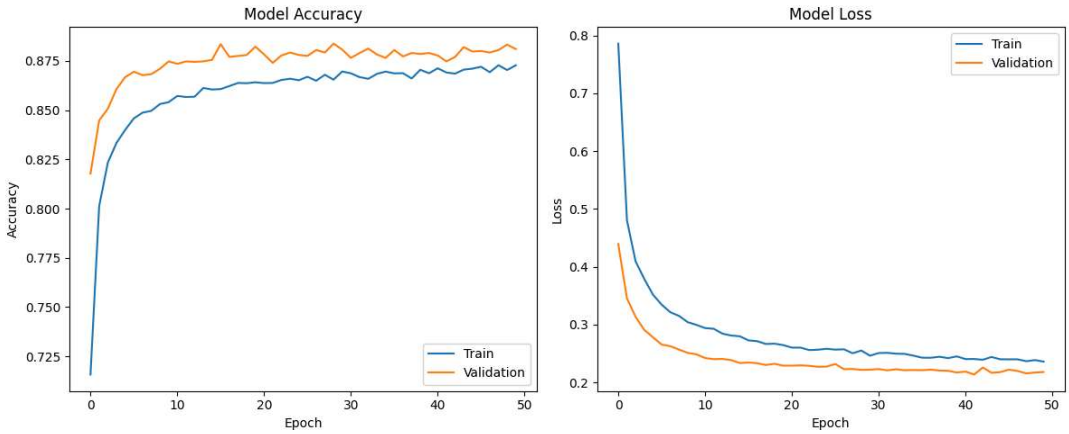
- All languages spike up positively around **100–150**, with **Marathi and Punjabi** slightly higher.
- This coefficient generally reflects the **overall tilt or shape** of the spectrum — the high values suggest dominant low frequencies across all languages.

3. MFCC Coefficients 2–12

- These coefficients capture **formant structure** and finer phonetic details**.**
- Most languages **hover close to zero** from coefficient 3 onward, but:
    - **Kannada** again stands out with a sharp dip around coefficient 2 and coefficient 4 (possible unique articulatory feature)**.**
    - **Marathi and Telugu** are relatively higher and smoother in this range**.**

The language classification model performs quite well, achieving an overall accuracy of **88%.** Both the training and validation accuracy curves show consistent improvement, with the validation accuracy stabilizing just under 89%**.** The loss curves also show a steady decrease, which indicates good learning and **no major signs of overfitting.**





Classification Insights

Languages like **Hindi, Tamil, Telugu, Malayalam, and Marathi** are classified with very high precision, recall, and F1-scores—typically above 97%—demonstrating the model's strong ability to distinguish these languages.

Gujarati shows a noticeable gap between precision and recall. While recall is high (meaning most actual Gujarati samples are correctly identified), precision is relatively low, indicating that samples from other languages—especially **Punjabi**—are often misclassified as Gujarati.

Punjabi is where the model struggles the most. It has a **very low recall of just 13%**, meaning the model rarely correctly identifies Punjabi samples. Most of them are being misclassified, particularly as Gujarati, as shown in the confusion matrix. This could suggest overlap in the acoustic features (MFCCs) between Punjabi and Gujarati, or possibly a lack of distinctive training data for Punjabi.

MFCC Feature Analysis

Looking at the MFCC variance plot, the **first coefficient (index 0)** holds the most variance across all languages, which is expected since early MFCCs usually capture overall energy and spectral shape. **Marathi** stands out with a much higher variance in the first coefficient, indicating greater variability or signal strength.

The mean MFCC plot shows that while there are some subtle differences in the average values of coefficients for each language, the patterns are largely similar beyond the first few coefficients. This suggests that the first few MFCCs carry the most distinctive information for language classification.

Summary

- The model performs very well on most languages, especially Hindi, Tamil, Telugu, Malayalam, and Marathi.
- Gujarati suffers from low precision due to heavy misclassification of Punjabi samples.
- Punjabi is the most challenging language for the model, with very low recall and F1-score.
- MFCCs are effective for language classification, with the first few coefficients contributing most to the classification decisions.

https://github.com/shripadpate20/Speech_Enhancement_and_Comparative_Analysis_of_Indian_Languages

Reference:

1)Universal Speech Representation Learning with Speaker Aware Pre-Training

Authors: Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, Xiangzhan Yu

Published on: arXiv (2021)

2) microsoft/unispeech-sat-base-100h-libri-ft

3)https://blog.unrealspeech.com/unispeech-sat-universal-speech-representation-learning-with-speaker-aware-pre-training/

4)https://huggingface.co/speechbrain/sepformer_rescuespeech