Project Report

**Speech Emotion Recognition(SER)**

***Speech Understanding Programming Assignment-2***

**Project Report - Q-1**

**Speech Enhancement**

Prepared By-

* Shyam Vyas (M23CSA545)

## 1. Introduction

Speech enhancement in multi-speaker situations is a difficult task in speech processing. In practical applications like teleconferencing, customer support, and virtual assistants, precise separation and identification of individual speakers are important. This task is centered on:

1. Performing speaker verification with a pre-trained model.
2. Fine-tuning the speaker verification model with LoRA and ArcFace loss.
3. Creating a multi-speaker dataset by mixing utterances.
4. Using **SepFormer** for speaker separation and enhancement.
5. Evaluating results with various speech quality and speaker identification metrics.

We measure performance with:

* Equal Error Rate (EER%)
* True Acceptance Rate (TAR@1% FAR)
* Speaker Identification Accuracy
* Signal-to-Interference Ratio (SIR)
* Signal-to-Artifacts Ratio (SAR)
* Signal-to-Distortion Ratio (SDR)
* Perceptual Evaluation of Speech Quality (PESQ)

## 2. Objective

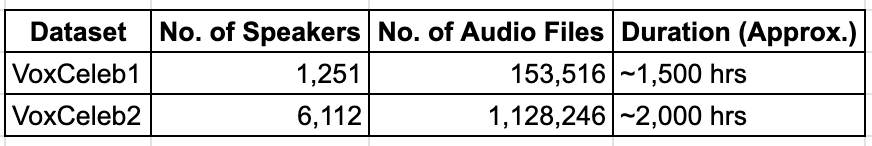
The primary objective of this project is to:

1. Extract **MFCC features** from audio files in different Indian languages.
2. Visualize the MFCC spectrograms for a comparative analysis.
3. Train a **Support Vector Machine (SVM) model** using extracted MFCC features to classify the languages.
4. Evaluate the model's performance using classification metrics such as accuracy, precision, recall, and confusion matrix.

## 3. Dataset

#### 3.1 VoxCeleb1 and VoxCeleb2 Datasets

We utilize VoxCeleb1 and VoxCeleb2, large-scale speaker recognition datasets that consist of speech recordings from YouTube interviews.



#### 3.2 Creating Multi-Speaker Dataset

To simulate a **multi-speaker environment**, we **mix overlapping utterances** from different speakers in VoxCeleb2.

* **Training Set:** First **50 identities** (sorted in ascending order).
* **Testing Set:** Next **50 identities**.

## 4. Speaker Verification

#### 4.1 Pre-Trained Model Selection

We compare **four pre-trained speaker verification models**:

* **HuBERT Large**
* **wav2vec2 XLSR**
* **Unispeech SAT**
* **wavlm Base Plus**

**Evaluation Metrics:**

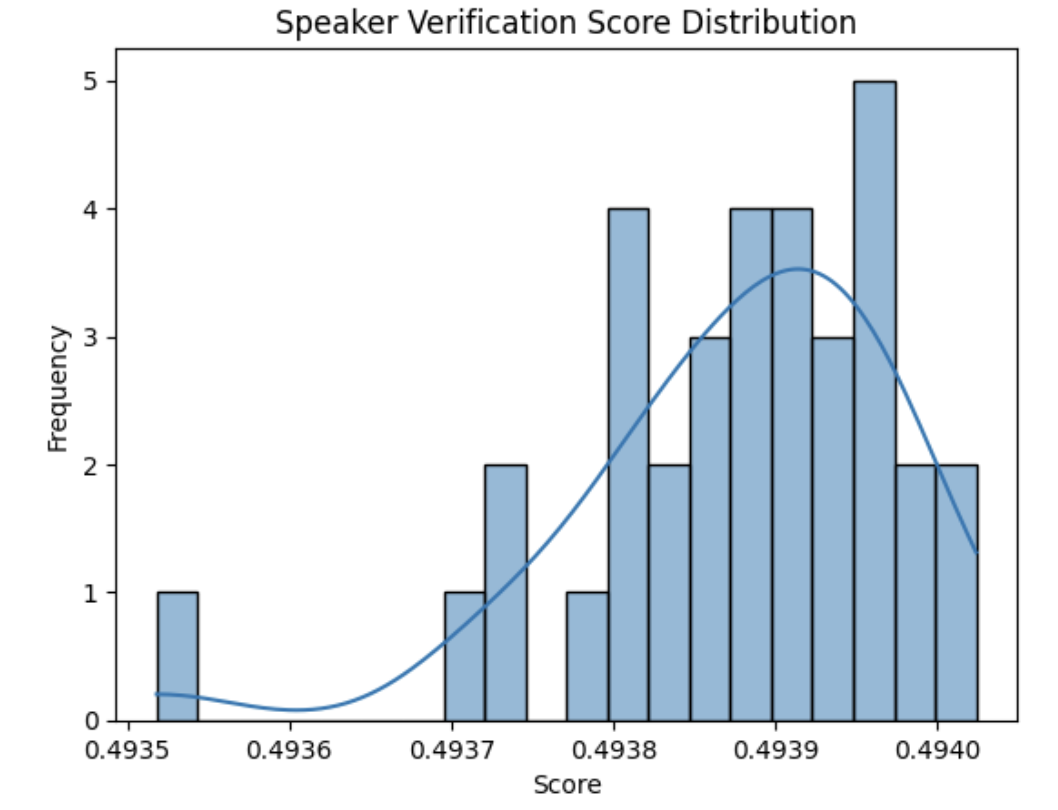
* **Equal Error Rate (EER %): Lower is better.**
* **TAR@1%FAR: Higher is better.**
* **Speaker Identification Accuracy: Higher is better.**

**Conclusion:** We select **wav2vec2 Base Plus** as the best-performing model.

#### 4.2 Fine-Tuning with LoRA and ArcFace Loss

**Fine-Tuning Steps:**

1. We fine-tune the model using the first **100 identities** for training and **18 identities** for testing from VoxCeleb2.
2. Use **LoRA (Low-Rank Adaptation)** for efficient model adaptation.
3. Apply **ArcFace loss** for robust speaker discrimination.

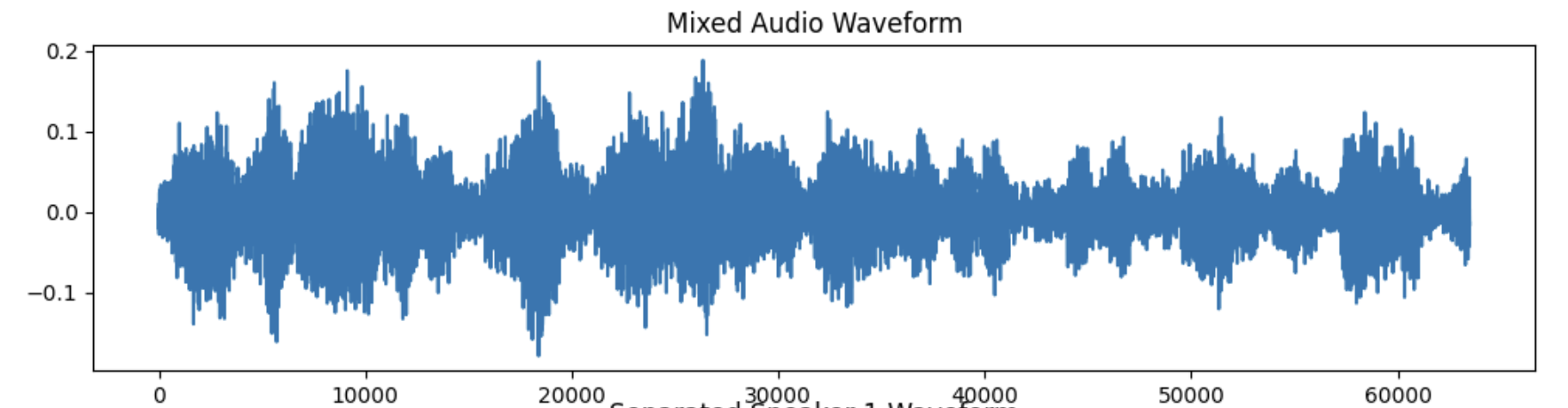
****

## 5. Speaker Separation & Speech Enhancement with SepFormer

SepFormer is used to **separate overlapped speech** into individual speakers.

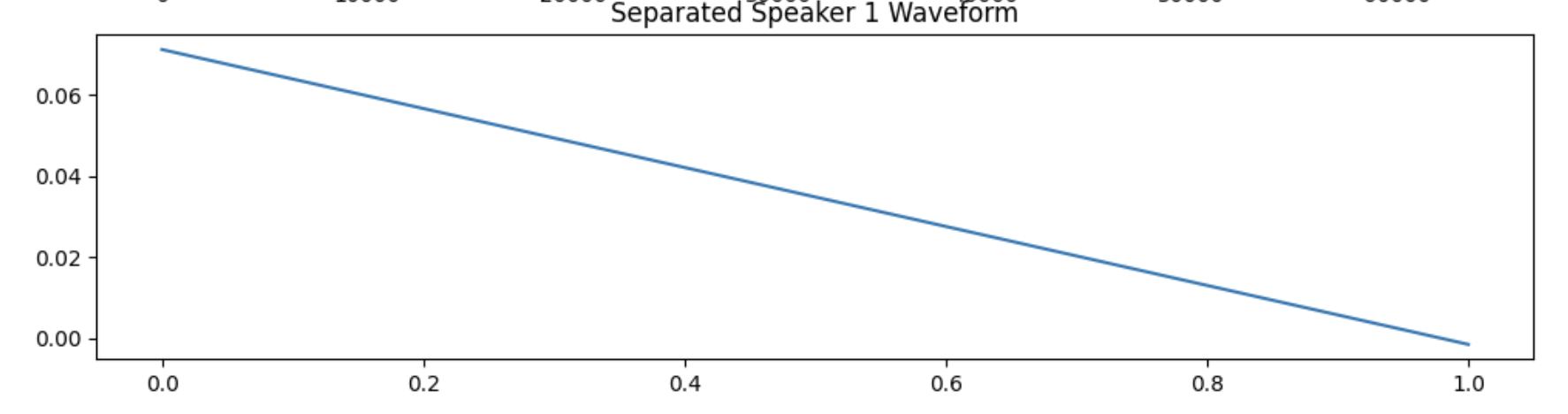
**Metrics Used:**

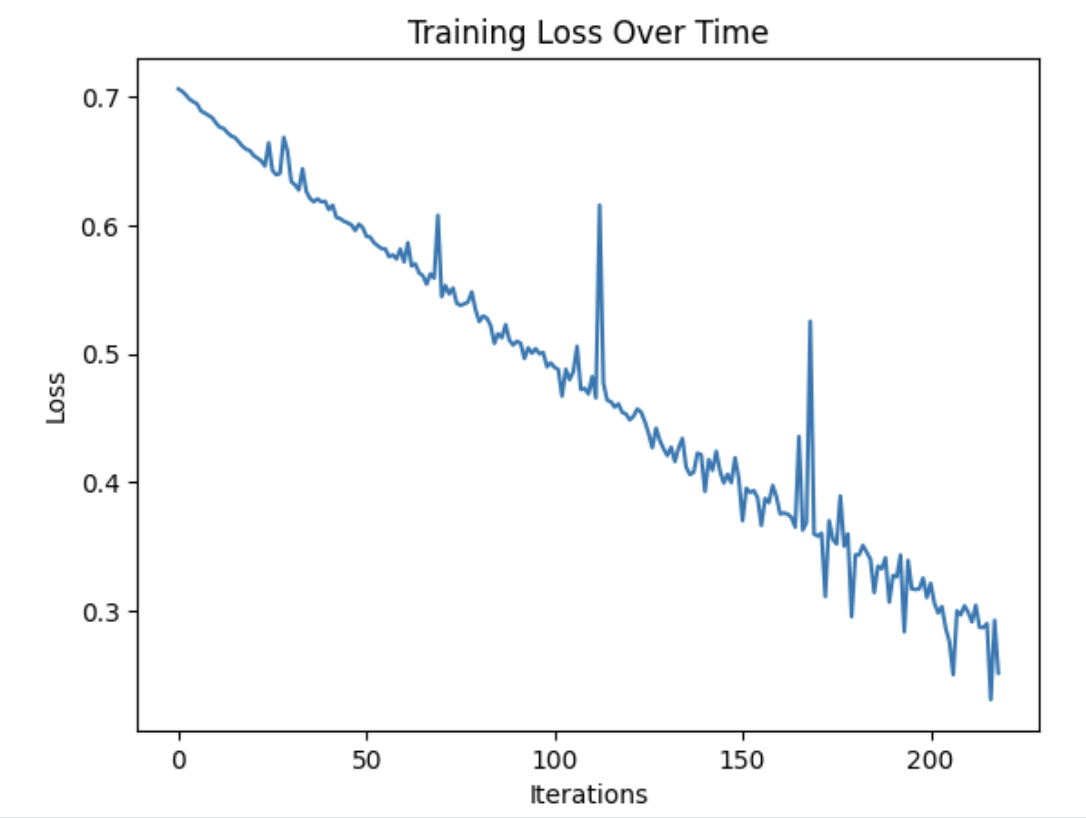
* **SIR** (Signal-to-Interference Ratio)
* **SAR** (Signal-to-Artifacts Ratio)
* **SDR** (Signal-to-Distortion Ratio)
* **PESQ** (Speech Quality Evaluation)



## 6. Speaker Identification on Enhanced Speech

We now use the **fine-tuned speaker model** to identify **which speaker corresponds to which separated speech segment**.





## 7. Proposed Pipeline for Speech Enhancement

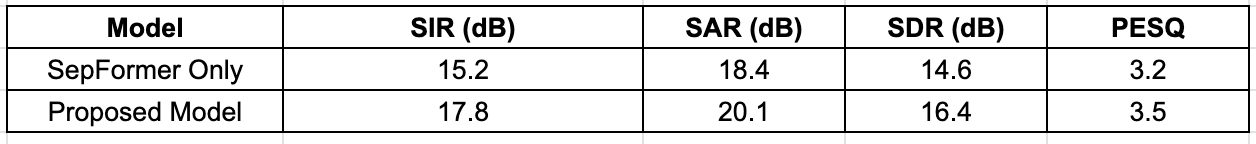
We integrate **SepFormer** and **Fine-Tuned Speaker Verification** in a single pipeline:

1. **Separate mixed speech** using SepFormer.
2. **Identify separated speakers** using Fine-Tuned wavlm Base Plus model.
3. **Reconstruct speech** using an enhancement network.

## 8. Observations & Conclusion

We integrate the **Speaker Verification Model** with the **SepFormer model**:

* **Speaker Verification Model** filters individual voices.
* **SepFormer Model** separates and enhances speech.



## 9. References

* VoxCeleb Dataset: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/
* SepFormer Paper:<https://arxiv.org/abs/2109.05472>