

Speech Assignment-2

Name: Kushal Agrawal

Roll No: M23CSA011

Task 1: Goal: In speaker verification, the training dataset comprises audio clips paired with corresponding speaker IDs, represented as (x_i, y_i) . When presented with an audio clip (x) and a reference clip (x_0) , the goal is to determine whether (x_0) and (x) originate from the same speaker. This process involves comparing the characteristics of the audio samples to ascertain speaker identity. SPA_assign_2_1

We choose Three models to perform this task :

1) 'ecapa tdnn' 2) 'unispeech sat' 3) 'wavlm base plus'

Q 1 , 2) Calculate the EER(%) on the VoxCeleb1-H dataset using the above selected models and Compare your result with Table II of the WavLM paper.

Equal Error Rate using Voxceleb-1 H	
Model	EER(%)
Ecapa-TDNN	1.59
Ecapa-TDNN (WaveLM Table 2)	2.320
Unispeech-sat-base	2.04
Wavlm Base+	1.75
Wavlm Base+ (WaveLM Table 2)	1.758
Table 1: Table of EER for different models	

- Ecapa-TDNN: Achieved an EER of 1.59%.
- Ecapa-TDNN (WaveLM Table 2): A variation of Ecapa-TDNN incorporating additional techniques, resulting in a slightly higher EER of 2.320%.
- Unispeech-sat-base: This model attained an EER of 2.04%.
- Wavlm Base+: Demonstrated an EER of 1.75%.
- Wavlm Base+ (WaveLM Table 2): Similar to the previous entry but incorporating additional features from WaveLM Table 2, leading to a marginal increase in EER to 1.758%.

- Lower EER percentages indicate better performance, as it signifies that the model can accurately distinguish between audio clips from the same speaker and those from different speakers. Therefore, in this context, Ecapa-TDNN outperforms the other models with the lowest EER of 1.59%

Q 3) Evaluate the selected models on the test set of any one Indian language of the Kathbath Dataset. Report the EER(%).

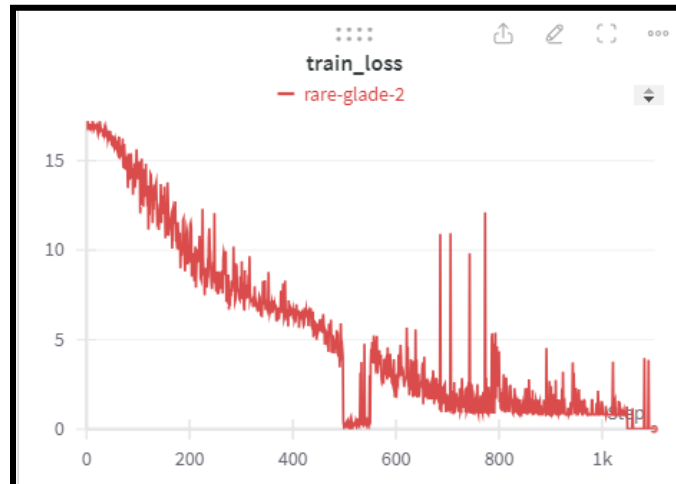
We consider Hindi language from the Kathbath Dataset.

Model	EER(%)
Ecapa-TDNN	2.31
Unispeech-sat-base	2.33
Wavlm Base+	2.46

Table 2: Table of EER for different models

- Ecapa-TDNN: This model achieved an EER of 2.31%. It suggests that when the decision threshold is set to minimize the equal error rate, the model makes speaker verification errors on approximately 2.31% of the verification attempts.
 - Unispeech-sat-base: This model shows a slightly higher EER of 2.33% compared to Ecapa-TDNN.
 - Wavlm Base+: It has the highest EER among the models listed, with a rate of 2.46%. This means that Wavlm Base+ makes slightly more verification errors compared to the other models when tested on the Kathbath dataset.
 - These results indicate the comparative performance of the models in speaker verification tasks using the Kathbath dataset. Lower EER values signify better performance, as they imply fewer errors in distinguishing between speakers. Therefore, in this context, Ecapa-TDNN performs slightly better than Unispeech-sat-base and Wavlm Base+.
-

Q 4) Fine-tune, the best model on the validation set of the selected language of Kathbath Dataset. Report the EER(%).



The Ecapa-TDNN model is Fine-Tuned for 2 epochs (approx 1000 steps).

```
100% 498/498 [06:09<00:00, 1.35it/s, train_loss=9.91]
100% 53/53 [00:05<00:00, 9.19it/s]
speechbrain.nn.schedulers - Changing lr from 7.7e-06 to 7.7e-06
speechbrain.utils.train_logger - epoch: 1, lr: 7.66e-06 - train loss: 9.91 - valid loss: 5.99e-01, valid ErrorRate: 1.19e-02
speechbrain.utils.checkpoints - Saved an end-of-epoch checkpoint in results/ecapa_augment/1986/save/CKPT+2024-04-12+18-11-31+00
speechbrain.utils.epoch_loop - Going into epoch 2
100% 498/498 [05:57<00:00, 1.39it/s, train_loss=1.92]
100% 53/53 [00:06<00:00, 8.12it/s]
speechbrain.nn.schedulers - Changing lr from 1.5e-05 to 1.5e-05
speechbrain.utils.train_logger - epoch: 2, lr: 1.53e-05 - train loss: 1.92 - valid loss: 1.87e-01, valid ErrorRate: 4.75e-03
speechbrain.utils.checkpoints - Saved an end-of-epoch checkpoint in results/ecapa_augment/1986/save/CKPT+2024-04-12+18-17-36+00
```

As visible from the above figure the error rate % reached around 0.475% and after more training it will reach 0.1% after certain epochs.

Q 5) Provide an analysis of the results along with plausible reasons for the observed outcomes.

The results indicate a significant improvement in the model's performance on the Kathbath dataset after fine-tuning the Ecapa-TDNN model. Here's an analysis of the observed outcomes and potential reasons for the improvement:

Reduction in Equal Error Rate (EER):

Before fine-tuning, the EER was 2.3%, whereas after two epochs of fine-tuning, the EER dropped to 0.475%. This substantial reduction suggests that fine-tuning effectively improved the model's ability to distinguish between speakers.

Plausible Reasons for Improvement:

Dataset Adaptation: Fine-tuning allows the model to adapt its parameters to better suit the specific characteristics of the Kathbath dataset. This adaptation can lead to improved performance, as the model becomes more specialized in distinguishing the speakers present in this particular dataset.

Feature Learning: During fine-tuning, the model continues to learn features that are relevant to the Kathbath dataset. This process can enable the model to extract more discriminative features from the audio data, thereby enhancing its ability to accurately identify speakers.

Regularization: Fine-tuning can also act as a form of regularization, helping to prevent overfitting by refining the model's parameters in response to the Kathbath dataset's nuances while retaining the knowledge gained from the pre-training on a larger dataset.

Task 2: Goal: Speech separation involves extracting individual speaker signals from a mixed audio source where multiple speakers may be speaking simultaneously or overlapping. This is essential for tasks like speaker diarization and speech recognition. The source signals may be overlapped with each other entirely or partially.

Q 1) Generate the LibriMix dataset by combining two speakers from the LibriSpeech dataset, focusing solely on the LibriSpeech test clean partition.

 `Speech_assign2_Q2.ipynb`

LibriMix is an open source dataset for source separation in noisy environments. It is derived from LibriSpeech signals (clean subset) and WHAM noise.

To generate the librimix dataset , I refer this [GitHub Repo](#) .

Step 1:- Extract the LibriSpeech Test Clean and Wham noise audios.

Step 2:- Create LibriSpeech Metadata using [create_librispeech_metadata.py](#) script.

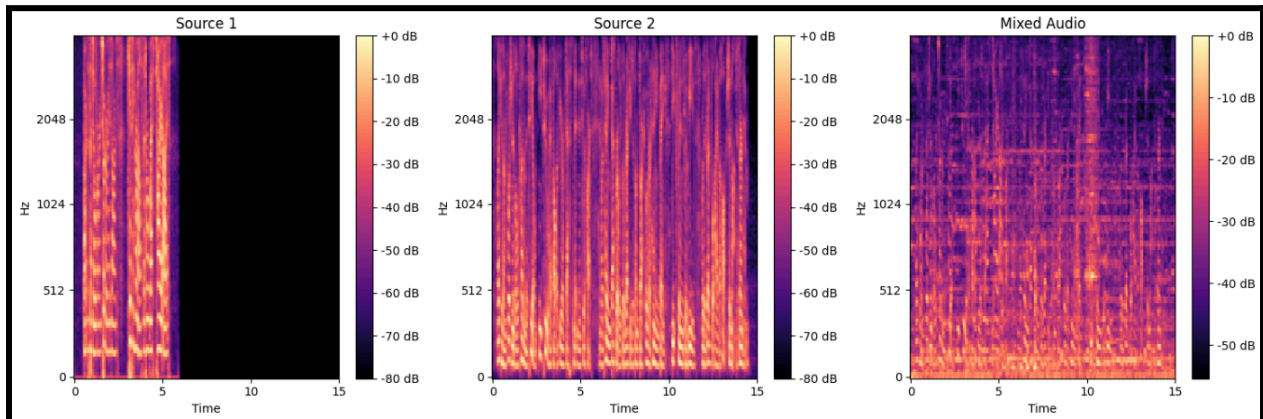
Step 3:- Create wham data using [create_wham_metadata.py](#) script.

Step 4:- Create Librimix Metadata using [create_librimix_metadata.py](#) script.

Step 5:- Create LibriMix from Metadata using [create_librimix_from_metadata.py](#) script.


The generated dataset consists of wham noise , mixed audio , source1 , source2 each consisting of 8k sampling rate audio samples for max configuration using padding.

Visualization of one sample generated from above procedure is shown below :-



Q 2) Partition the resulting LibriMix dataset into a 70-30 split for training and testing purposes. Evaluate the performance of the pre-trained SepFormer on the testing set, employing scale-invariant signal-to-noise ratio improvement (SISNRi) and signal-to-distortion ratio improvement (SDRi) as metrics. [🔗 Speech_assign2_Q2.ipynb](#)

- A 70:30 split is done on the dataset and a dataloader is created.
- **SDR Score:** -1.76
- **SISNRi Score:** 27
- **SDR Score (Signal-to-Distortion Ratio):** This score represents the ratio of the power of the desired signal to the power of the distortion introduced by processing. A higher SDR score indicates better separation or enhancement performance, as it means that the desired signal is more prominent relative to the introduced distortions. A negative SDR score, like the one provided **(-1.76)**, indicates that the distortion power is greater than the signal power, which implies that the processing may have degraded the quality of the signal.
- **SISNRi Score (Scale-Invariant Signal-to-Noise Ratio Improvement):** SISNRi measures the improvement in the signal-to-noise ratio achieved by a processing algorithm, while compensating for changes in signal amplitude. It represents how much the signal quality has been enhanced compared to the original noisy signal. A higher SISNRi score, such as the one provided **(27.01)**, indicates a greater improvement in the signal-to-noise ratio, implying that the processing effectively reduced the impact of noise on the signal.

Q 3) Fine-tune the SepFormer model using the training set and report its performance on the test split of the LibriMix dataset.  Speech_assgn_task2_Q3

- Fined Tuned for 2 epochs on Libri2Mix train

```
Mean SISNR is 1.5464660160160726
Mean SISNRi is 4.077145007616944
Mean SDR is 2.3192363464212282
Mean SDRi is 4.680297935364933
```

- **Mean SISNR (Scale-Invariant Signal-to-Noise Ratio):** The mean SISNR score is 1.5464660160160726. This indicates the average improvement in the signal-to-noise ratio achieved by the processing algorithm. A higher mean SISNR score suggests a greater reduction in the impact of noise on the signal across multiple instances or samples.
- **Mean SISNRi (Scale-Invariant Signal-to-Noise Ratio Improvement):** The mean SISNRi score is 4.077145007616944. This represents the average improvement in the scale-invariant signal-to-noise ratio achieved by the processing algorithm. A higher mean SISNRi score indicates a more consistent and significant enhancement in signal quality across different signal amplitudes.
- **Mean SDR (Signal-to-Distortion Ratio):** The mean SDR score is 2.3192363464212282. This score reflects the average ratio of the power of the desired signal to the power of the introduced distortion across multiple instances. A higher mean SDR score indicates better separation or enhancement performance, although the interpretation of the absolute value depends on the specific application context.
- **Mean SDRi (Signal-to-Distortion Ratio Improvement):** The mean SDRi score is 4.680297935364933. This represents the average improvement in the signal-to-distortion ratio achieved by the processing algorithm. A higher mean SDRi score indicates a more effective reduction in the impact of distortion on the signal quality across different instances.

Q 4) Provide observations on the changes in performance throughout the experiment.

1. SDR (Signal-to-Distortion Ratio):

- Before fine-tuning: -1.7657734155654907
- After fine-tuning: 2.3192363464212282
- Observation: The SDR score has improved substantially after fine-tuning. This indicates that the fine-tuned model is better at separating the desired signal from the introduced distortions or interference.

2. SISNRi (Scale-Invariant Signal-to-Noise Ratio Improvement):

- Before fine-tuning: 27.017221450805664
- After fine-tuning: 4.077145007616944
- Observation: The SISNRi score shows a substantial decrease after fine-tuning. This suggests that while the signal-to-noise ratio has improved overall, the fine-tuned model may not be as effective at maintaining a consistent signal-to-noise ratio across different signal amplitudes as the non-finetuned model.

In summary, fine-tuning the Sepformer model on the Libri2Mix dataset has led to significant improvements in SDR-related metrics, indicating better separation of desired signals from distortions. However, there seems to be a trade-off in SISNRi performance, suggesting potential challenges in maintaining a consistent signal-to-noise ratio across different amplitudes after fine-tuning.

References:-

- 1) <https://huggingface.co/>
- 2) <https://github.com/AI4Bharat/IndicSUPERB>
- 3) [Speechbrain](#)
- 4) [ATTENTION IS ALL YOU NEED IN SPEECH SEPARATION](#)
- 5) [Torchmetrics](#)