Speech Understanding Programming Assignment-2

Git Repository for the Assignment:

https://github.com/parasharharsh16/Speech-Understanding-PA2

wandb-link: https://wandb.ai/parasharharsh16/SU-assignment-finetunning-hubert?nw=nwuserparasharharsh16

Question-1

A. Calculated EER% on given 3 model on VOxCeleb1-H dataset
   1. wav2vec2_xlsr_finetune
   2. hubert_large_finetune
   3. wavlm_large_finetune

Model Name	EER%	EER Threshold
wav2vec2_xlsr_finetune	36.35%	0.9816
hubert_large_finetune	44.94%	0.9906
wavlm_large_finetune	44.36%	0.9819

B. Comparison of the result above with Table II of the WavLM paper.
   The EER % given in table II of WavLm paper for HuBERT Large is **1.342** and for WavLM Large is **0.986.**
   This EER % is significantly higher that what is reported in section A in my report. There can be many reasons to these results, but according to me below two are major reason.
   1. The number of data items taken for EER calculations (I have taken 1% of the data to calculate EER due to system resource constraints)
   2. The normalization of waveform can cause the discerption in the embeddings generated by the model

C. Testing the model and calculating EER for different combination of language with “Hindi”

Model	Lang	EER	EER Threshold
wav2vec2_xlsr_finetune	Hindi, Punjabi	29%	0.977
wav2vec2_xlsr_finetune	Hindi, Tamil	52%	0.979
wav2vec2_xlsr_finetune	Hindi, Sanskrit	43%	0.975
hubert_large_finetune	Hindi, Punjabi	49%	0.991
hubert_large_finetune	Hindi, Tamil	52%	0.988
hubert_large_finetune	Hindi, Sanskrit	42%	0.993
wavlm_large_finetune	Hindi, Punjabi	58%	0.923
wavlm_large_finetune	Hindi, Tamil	36%	0.924
wavlm_large_finetune	Hindi, Sanskrit	71%	0.936

D. Fine tuning result of best model on Kathbath Dataset

   The best performing model above, as overserved was hubert_large, and performing the finetuning of the model is done using 10% of Kathbath data, due to resource constraints

   The model has been trained with Hindi and Punjabi language combinations and then EER calculated on the test data of that language combination.

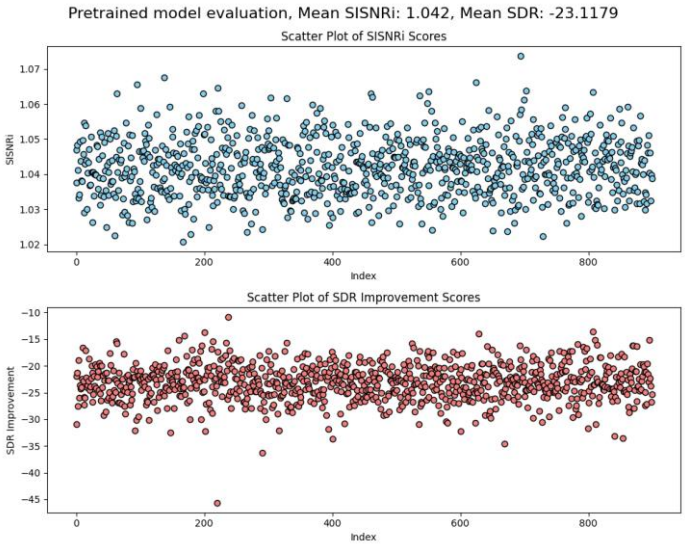
Speech Understanding Programming Assignment-2

Lang	EER	EER Threshold
hindi, punjabi	29.17%	0.9206

Question-2

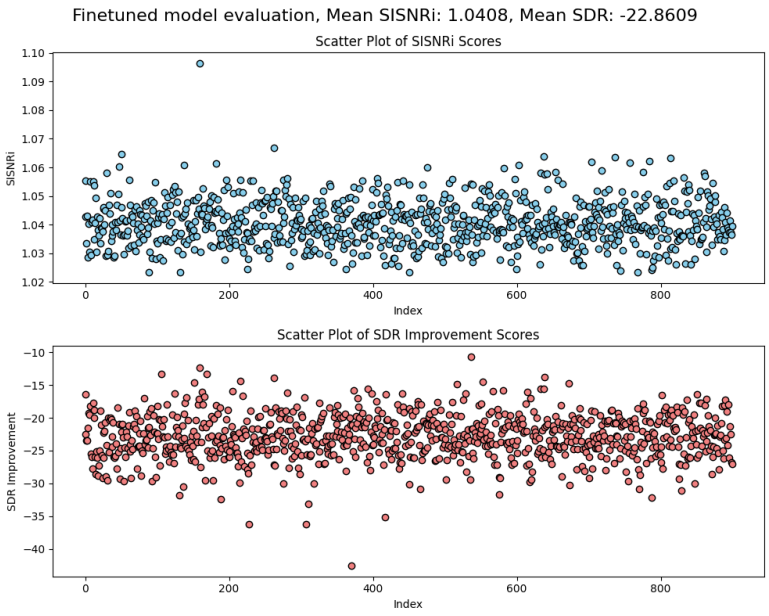
A. The scale-invariant signal-to-noise ratio improvement (SISNRi) and signal-to-distortion ratio improvement (SDRi) for the sepformer model which was calculated on Librimix test-clean dataset (total 900 records i.e. 30% of total dataset)

Average SISNRi	1.042
Average SDRi	-23.1179



B. Model is being finetuned on 20 % of train data, due to system constraints and evaluation of finetuned model is given below:

Average SISNRi	1.0408
Average SDRi	-22.8609



Speech Understanding Programming Assignment-2

References:

- https://github.com/JorisCos/LibriMix
- https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification
- https://github.com/AI4Bharat/IndicSUPERB
- https://huggingface.co/speechbrain/sepformer-whamr