Total Marks: 40

Speech Understanding
Programming Assignment 2

Due date: 31/03/2024

Assignment Guidelines

1. Any kind of plagiarism is not accepted. We will strictly follow institute policies for plagiarism.
2. Recommended programming languages: Python + PyTorch.
3. You may use any external libraries or GitHub codes. However, the evaluation will test your knowledge of the algorithm and the choice of hyperparameters. Do cite the libraries/codes.
4. Please use wandb to log the metrics.
5. **A single zip file containing the report, codes and readme if required. The zip file should be named Rollno_PA2.zip.A Gradio link for the trained model to evaluate the trained model on the browser.**
6. **You must also provide a link to your GitHub code. Follow good software engineering practices and set up a GitHub repo for the project on Day 1. Please do not write all code on your local machine and push everything to Git Hub on the last day. The commits in GitHub should reflect how the code has evolved during the assignment. We will check regular commits while grading the assignments.**

Question 1. Speaker Verification

Goal: In speaker verification, the training dataset consists of audio clips paired with speaker IDs, denoted as $(D = (x_i, y_i))$. Given an audio clip (x) and a reference clip (x_0), the objective is to ascertain whether (x_0) and (x) belong to the same speaker.

Tasks: —- Choose three pre-trained models from the list: 'ecapa_tdnn', 'hubert_large', 'wav2vec2_xlsr', 'unispeech_sat', 'wavlm_base_plus', 'wavlm_large' trained on the VoxCeleb1 dataset. You can find the pre-trained models on this link.

—- Calculate the EER(%) on the VoxCeleb1-H dataset using the above selected models. You can get the dataset from here. [3 Marks]

—- Compare your result with Table II of the WavLM paper. [2 Marks]

—- Evaluate the selected models on the test set of any one Indian language of the Kathbath Dataset. Report the EER(%). [2 Marks]

—- Fine-tune, the best model on the validation set of the selected language of Kathbath Dataset. Report the EER(%). [10 Marks]

—- Provide an analysis of the results along with plausible reasons for the observed outcomes. [5 Marks]

Question 2. Source Separation

Goal: The goal of speech separation is to estimate individual speaker signals from their mixture, where the source signals may be overlapped with each other entirely or partially.

Tasks: —- Generate the LibriMix dataset by combining two speakers from the LibriSpeech dataset, focusing solely on the LibriSpeech_test_clean partition. Take help from this GitHub repo. [3 Marks]

—- 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. For metric computation, consult the provided paper and utilize the code from torchmetrics. [3 Marks]

—- Fine-tune the SepFormer model using the training set and report its performance on the test split of the LibriMix dataset. [10 Marks]

—- Provide observations on the changes in performance throughout the experiment. [4 Marks]