Team Name: Team-OPTIMAL

Contact Person: Prof. M. Tanveer

Email ID: mtanveer@iiti.ac.in

Model Description and Training Details

Model Overview:

AUREXA-SE (Audio–Visual Unified Representation Exchange Architecture with Cross-Attention and Squeezeformer for Speech Enhancement) is a bimodal speech enhancement framework designed for the COG-MHEAR AVSE Challenge. It integrates a **U-Net–based audio encoder** that processes raw waveforms and a **Swin Transformer V2 visual encoder** that extracts spatially rich features from facial video frames. These modality-specific embeddings are fused using a **bi-directional cross-attention mechanism**, enabling deep contextual interaction. Temporal dependencies are modelled using **Squeezeformer blocks**, and the final enhanced representation is decoded into clean speech using a **U-Net–style waveform decoder**.

Training Specifications:

- Memory Footprint: During training, the model utilized 146 GB of shared RAM on an NVIDIA RTX A4500 GPU
- Hardware Specifications: Training and inference were conducted on a single NVIDIA RTX A4500 GPU equipped with 146 GB of shared RAM
- No data augmentation techniques were applied during training
- No. of trainable parameters is ~54.5M, with 217.859 MB total estimated model parameters size

Training Process Details:

- **Training Duration**: ~20 epochs with early stopping based on validation SI-SDR, STOI, PESQ. Averages over all scenes: PESQ: 1.325 STOI: 0.514 SI-SDR: -4.312
- Each epoch required about 2 hours 30 minutes using GPU acceleration
- **Number of Training Steps**: The training process had a total of 344,680 steps.
- **Input Modalities**: Raw binaural audio (converted to mono) and 75-frame RGB video clips (112×112 resolution).
- **Loss Function**: Combination of time-domain reconstruction loss and perceptual metrics (MSE, SI-SNR, STOI and PESQ).
- **Optimizer**: Adam with learning rate scheduling.
- **Batch Size**: Tuned based on GPU memory; typically, 2-3.

Reproducibility:

Link to GitHub repo - https://github.com/mtanveer1/AVSEC-4-Challenge-2025

System Constraints and Requirements:

- Limitations: There are no known limitations or constraints specific to this system.
- Requirements: Fundamental data science libraries like numpy, pandas.
- NVIDIA Cuda framework for GPU accelerated training and interference, PyTorch Lightning framework.

  Exact version specific requirements mentioned in Github repository