We propose a time-domain audio-visual speech enhancement (SE) model based on transformers for the 1st COG-MHEAR audio-visual speech enhancement challenge. The model exploits noisy speech, target speakers face and pose-invariant landmark flow features to estimate clean speech in time domain.

Audio feature extraction: The time domain audio signals are encoded and decoded using 1-D convolutional and transpose convolutional neural networks. The encoded audio signal is fed to a cross-attention encoder module. The cross attention modules present in the encoder comprise 4 heads and 16 dimensions each.

Visual feature extraction: The visual feature extraction consists of RESNET-18 to extract facial attribute features given a cropped face region. The extracted facial features are upsampled to match the video sampling rate. The upsampled facial attribute feature is combined with pose-invariant landmark flow features to generate final visual features.

Multimodal fusion: The upsampled visual features and encoded audio features are concatenated and fed to a series of three transformer modules. The self attention head present in each transformer module consists of 4 heads and 16 dimensions per head. The processed latent space is then fed to a decoder module that maps the latent space to the output dimension after applying cross attention. The cross attention modules present in the encoder comprise 4 heads and 16 dimensions each.

The model is trained and fine tuned using AVSE train and dev set respectively. Table 1 demonstrated the overview results for objective evaluation on the dev set. It can be seen that for all objective measures the proposed AV outperforms baseline. The test set outputs are processed using Audacity Telephone Filter EQ to remove artefacts present.

| | PESQ | STOI | SI-SDR | SI-SNR |
|---|---|---|---|---|
| Noisy | 1.1549 | 0.6396 | -5.1008 | -4.6886 |
| Baseline | 1.3065 | 0.6749 | 2.4769 | 2.4789 |
| Proposed | 1.6988 | 0.8320 | 9.8521 | 9.8637 |