A TIME-DOMAIN AUDIO-VISUAL SPEECH ENHANCEMENT MODEL FOR 3rd COG-MHEAR AVSE CHALLENGE

Mandar Gogate, Kia Dashtipour, Amir Hussain

Edinburgh Napier University, Edinburgh, UK

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 pretrained AV Hubert to extract video features given a cropped lip 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 eval set.

	PESQ	STOI	SI-SDR
Noisy	1.4672	0.6103	-5.4942
Proposed	2.58487	0.8512	10.9021