ROBUST AUDIO-VISUAL SPEECH ENHANCEMENT IN THE WAVEFORM DOMAIN FOR 1ST COG-MHEAR AVSE CHALLENGE

1. PROPOSED APPROACH

We propose a time-domain audio-visual speech enhancement (SE) model based on transformers [1] as depicted in Fig. 1 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.

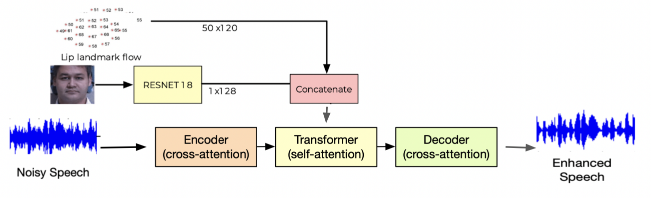
Lip landmark flow
50 x120
RESNET 18
1x128
Concatenate
Noisy Speech
Encoder (cross-attention)
Transformer (self-attention)
Decoder (cross-attention)
Enhanced Speech

Fig. 1. Proposed Framework

Audio feature extraction: The time domain audio signals are encoded and decoded using 1-D convolutional and transpose convolutional neural network as proposed in [2]. The encoded audio signal is fed to a cross-attention encoder module as shown in Fig. 1. The cross attention modules present in the encoder comprise of 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 dimension 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 of 4 heads and 16 dimensions each.

2. EXPERIMENTAL RESULTS

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. We also evaluated subjective listening quality on test set using a non-intrusive perceptual objective speech quality metric - DNS-MOS [3]. Table 2 presents the DNS MOS results for eval set.

Table 1. Objective evaluation on dev set

	PESQ	STOI	SI-SDR	SI-SNR
Noisy	1.154	0.639	-5.100	-4.688
Baseline	1.306	0.674	2.476	2.478
Proposed AV	1.698	0.841	9.852	9.863
Oracle IBM	1.974	0.907	12.539	12.710

Table 2. DNS MOS [3] results on eval set

	OVRL	SIG	BAK
Noisy	1.9150	2.7545	2.0873
Baseline	1.9802	2.5880	2.7854
Proposed AV	2.3602	2.8421	3.2860

3. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[3] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 886–890.