# AVFORMER: TRANSFORMER BASED AUDIO-VISUAL SPEECH SEPARATION FOR 2ND COG-MHEAR AVSE CHALLENGE

*Mandar Gogate, Kia Dashtipour, Amir Hussain*

Edinburgh Napier University, UK
m.gogate@napier.ac.uk

## 1. PROPOSED APPROACH

We propose a time-domain audio-visual speech separation (SE) model based on Transformer [1] as depicted in Fig. 1 for the $2^{nd}$ COG-MHEAR audio-visual speech enhancement challenge. The model exploits noisy speech, and target speakers lips to supress unwanted background noise and enhance the target speech.

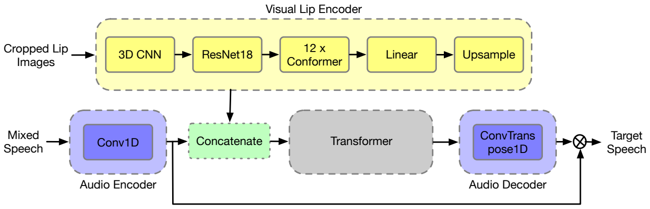

**Fig. 1**. Proposed Framework

**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 convolutional layer had 16 filters and 8 stride size.

**Visual feature extraction**:- The visual feature extraction stage of the pipeline consist of 3D convolutional layer with filter size of 5 × 7 × 7 and stride of 1 × 2 × 2, followed by ResNet-18. The ResNet-18 is followed by conformer for temporal modelling [3]. The extracted visual features are upsampled to match the video sampling rate.

**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.

**Table 1**. Objective evaluation on eval set (leaderboard)

| | PESQ | STOI | SI-SDR |
|---|---|---|---|
| Noisy | 1.136 | 0.441 | -5.073 |
| Baseline | 1.414 | 0.556 | 3.667 |
| Proposed | 2.110 | 0.781 | 12.355 |

**Table 2**. System parameters

| Parameter | |
|---|---|
| No. of parameters | 22.1 M |
| Training epochs | 100 |
| Latency (M1 Macbook Pro) | 0.5 sec processing time for 1 sec of video |
| Training hardware | 2 x NVIDIA RTX A6000 |
| Time per epoch | 60 min |
| Total training steps | 200k |
| Batch size | 4 per GPU |
| Training Time | 5 days |
| Total epochs | 100 |
| Optimiser | Adam |
| Learning rate | 0.003 |
| Learning rate scheduler | Multiply learning rate by 0.8 if validation loss stops decreasing for two epochs |
| DNN Framework | PyTorch |

## 2. EXPERIMENTAL RESULTS

Table 1 demonstrated the overview results for objective evaluation on the AVSEC 2 leaderboard eval set. It can be seen that for all objective measures the proposed AV outperforms baseline. Table 2 presents the system parameters.

## 3. REFERENCES

[1] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[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] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7613–7617.