Multi-Model Dual-Transformer Network-based Audio-Visual Speech Enhancement for 3rd COG-MHEAR AVSE Challenge

Fazal E Wahab[a], Nasir Saleem[b], Amir Hussain[c]

[a] *University of Science and Technology of China, , Hefei, China*
[b] *Gomal University, D.I.Khan, Pakistan*
[c] *Edinburgh Napier University, Edinburgh, UK*

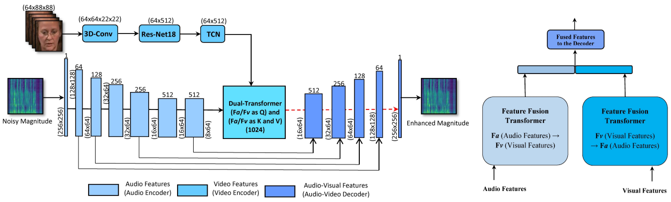
(64x88x88)
(64x64x22x22)
3D-Conv
(64x512)
Res-Net18
(64x512)
TCN
Noisy Magnitude
1
64
128
256
256
512
512
(256x256)
(128x128)
(64x64)
(32x64)
(32x64)
(16x64)
(16x64)
(8x64)
Dual-Transformer (Fa/Fv as Q) and (Fa/Fv as K and V) (1024)
512
256
128
64
1
(16x64)
(32x64)
(64x64)
(128x128)
(256x256)
Enhanced Magnitude
Audio Features (Audio Encoder)
Video Features (Video Encoder)
Audio-Visual Features (Audio-Video Decoder)
Fused Features to the Decoder
Feature Fusion Transformer
Fa (Audio Features) → Fv (Visual Features)
Feature Fusion Transformer
Fv (Visual Features) → Fa (Audio Features)
Audio Features
Visual Features

Figure 1: Architecture of the proposed AVSE. The baseline model is selected with a Dual-Transformer block for self-supervised feature learning and fusion.

1. Proposed Multi-Model AVSE Approach

We propose a multi-model dual-transformer that uses the attention mechanism to capture correlations between features for audio-visual speech enhancement. The transformer independently processes the audio and visual features before fusing them in a self-supervised manner. The proposed model leverages both the noisy speech input and the visual cues from the target speaker's lip movements to effectively reduce unwanted background noise and enhance the intelligibility and quality of the target speech. Figure 1 demonstrates the proposed AVSE model for the 3rd COG-MHEAR AVSE Challenge.

1.1. Audio and Visual Encoder

The audio feature extraction involves using a U-net-inspired network modified for AVSE Hussain et al. (2024), including encoder and decoder blocks. The audio encoder takes the STFT magnitude of noisy speech as input with ($F \times T$) dimensions. The input is processed by two convolutional layers with a filter size of 4 and a stride of 2. Following this, the reduced features map passes through three convolutional blocks. The visual feature extraction starts with a 3D convolutional layer with a (5×7×7) filter size and a (1×2×2) stride, followed by ResNet-18. Subsequently, the features from the residual network are led into a Temporal Convolutional Network (TCN) Luo and Mesgarani (2019). The inputs consist of a sequence of cropped lip images, each sized ($M \times 88 \times 88$). From these lip images, the visual feature encoder generates a 512-dimensional vector for each.

Table 1: System parameters

Parameter Type	Observations
Number of Parameters (Millions)	44.68
Training Epochs	100
RTF	0.9751 Seconds (On CPU) 0.0106 Seconds (On GPU)
Model-Size (MB)	170.62
Training Hardware	NVIDIA GeForce GTX1660Ti Intel® Core™ i5-4300 CPU@1.90 GHz
Time Per Epoch	16 Minutes
Total Training Steps	136K
Batch Size	5
Training Time	27 Hours
Total Epochs	100
Optimiser	Adam
MACs (G)	227.95
Learning Rate	0.003
DNN Framework	PyTorch

1.2. Multi-Model Dual-Transformer Block

Figure 1 illustrates that the attention mechanism in the Transformer model Subakan et al. (2021) allows for a focused consideration of the relationship between audio-visual features. This mechanism greatly aids in facilitating the complete fusion of audio and video features. This model uses a 4-layer transformer encoder to achieve this multi-model feature fusion. The attention mechanism within the transformer is expressed as Q (query vector) and KV (queried vector). The visual features are adjusted in dimension to align with audio features (512×512) and fed to the multi-model dual-transformer block. Next, the block uses one of the features as Q and the other as KV. These features are then fused. The fused audio-visual features are fed to the decoder. The resultant predicted mask is then multiplied element-wise with the input magnitude spectrogram, yielding the output spectrogram. Table 1 gives the system parameters.

References

Hussain, T., Dashtipour, K., Tsao, Y., Hussain, A., 2024. Audio-visual speech enhancement in noisy environments via emotion-based contextual cues. arXiv preprint arXiv:2402.16394 .

Luo, Y., Mesgarani, N., 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing 27, 1256–1266.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J., 2021. Attention is all you need in speech separation, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 21–25.