AV-TFLocoformer: A Locally Convolutional Transformer for Robust Audio-Visual Speech Enhancement

Aquib Raza Shafique Ahmed

1 Signal Preparation

The speech signal is transformed to the time–frequency domain via a 512-point short-time Fourier transform (STFT) using a 25 ms Hann window with 50 % frame overlap. The full complex spectrum—concatenated real and imaginary parts is retained, allowing a single complex valued mask to restore both magnitude and phase while halving memory usage, yielding $\mathbf{X} \in \mathbb{R}^{2\times T\times 257}$. For the visual stream, mouth-region frames are centre-cropped to 96×96 pixels and temporally down-sampled to 25 fps.

2 Model Architecture

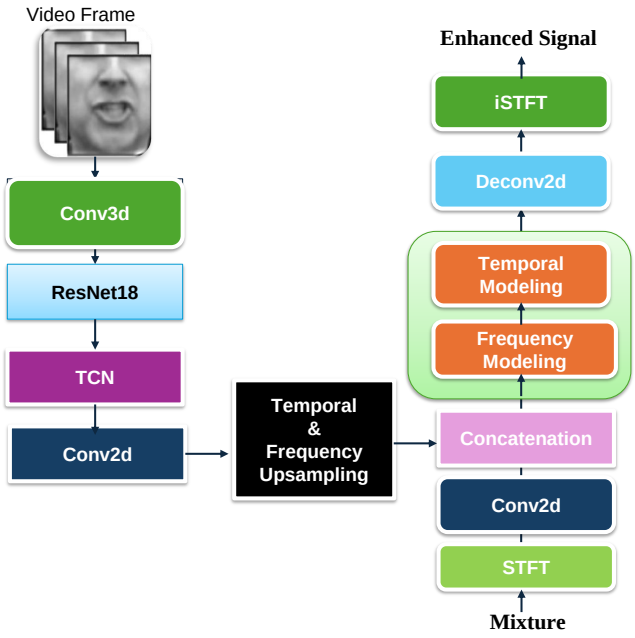

Figure 1: Overview of the proposed AV-TFLocoformer architecture.

Audio branch. A 3×3 Conv2D layer (input $2\to 64$ channels, stride 1, padding 1) followed by GroupNorm_{16} and PReLU produces $\mathbf{X}_a \in \mathbb{R}^{64\times T\times 257}$.

Visual branch. *(a) 3-D front-end:* A $1\to 32$ Conv3D layer (kernel $5\times 7\times 7$; stride $1\times 2\times 2$; padding $2\times 3\times 3$) is followed by $1\times 2\times 2$ max-pooling, preserving the temporal rate while reducing the spatial resolution to 12×12 pixels. *(b) Frame-wise ResNet:* Three residual blocks (32–32–32) refine spatial details. *(c) Temporal CNN:* A five-layer depth-wise separable TCN (kernel 3; dilation factors 1–2–4–8–1; dropout 0.1) captures mouth-motion dynamics and outputs $\mathbb{R}^{64\times T'}$. *(d) 2-D projection:* The 1-D feature stream is reshaped to $T'\times 1\times 64$; a 1×1 Conv2D ($64\to 32$) yields a *spectrogram-like* map $\mathbf{X}_v \in \mathbb{R}^{32\times T'\times F'}$. *(e) Temporal and frequency up-sampling:* Nearest-neighbour interpolation in time and bilinear interpolation in frequency resize $\mathbf{X}_v$ to match the audio grid $(T, 257)$. Both modalities are therefore represented as 2-D feature volumes, ready for fusion.

Early fusion. Channels are concatenated to obtain

$$\mathbf{X}_c = \left[\mathbf{X}_a;\, \mathbf{X}_v\right] \in \mathbb{R}^{96\times T\times 257}.$$

TF-Locoformer backbone. Two TF-Locoformer blocks [1] adopt a macaron structure: depth-wise dilated Conv2D $(5\times 5)\to$ eight-head multi-head self-attention (along **frequency**) with rotary position embedding (RoPE) $\to$ a second Conv2D, after which the trio is repeated along **time**. LayerNorm and a residual scaling factor of 0.5 maintain numerical stability in FP16 for $T \le 750$. These blocks account for approximately 1.2 M of the total 1.5 M parameters.

Mask head and reconstruction. A 3×3 Conv2D layer ($96\to 2$) produces the complex mask $\hat{\mathbf{M}}_c$. The enhanced spectrum is obtained by element-wise multiplication $\hat{\mathbf{S}} = \hat{\mathbf{M}}_c \odot \mathbf{X}$; an inverse STFT with overlap-add synthesis reconstructs the time-domain waveform.

Computational complexity. The complete network contains 1.5 M parameters and requires approximately 4.9 GFLOPs per second of audio. On an RTX-2080 Ti GPU (FP16, batch = 1) the real-time factor is about 0.9.

3 Training Configuration

The model is trained to minimise the negative scale-invariant SNR (SI-SNR) on 1-s waveform segments. Adam optimisation (learning rate 3×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$) is run for up to 25 epochs with early stopping (patience = 5). Training uses a mini-batch size of 2 on a single RTX-2080 Ti GPU; PyTorch AMP provides mixed-precision computation, and gradients are clipped to a global norm of 5.

References

[1] T. Saijo, K. Yatabe, and Y. Oikawa, "TF-Locoformer: Locally Convolutional Transformer for Speech Enhancement," *Proc. ICASSP*, 2024.