AV - LocoFiLM: Audio-Visual Speech Enhancement with Early FiLM Fusion and Local–Global Transformers

Shafique Ahmed Jen-Cheng Hou Yu Tsao

1 System Architecture

Pre-processing. 16 kHz waveforms are transformed to *complex* spectrograms by a 512-point STFT with 50 % overlap (256-sample hop). The four resulting channels—real, imaginary, magnitude and wrapped phase—are stacked into $\mathbf{X} \in \mathbb{R}^{4\times T\times F}$ ($F = 257$). Video is cropped to a 96×96 mouth ROI at 25 fps.

Fusion stem.

- *Audio branch.* A 3×3 Conv2D (4→128) with GroupNorm_{32} and PReLU converts $\mathbf{X}$ to $\mathbf{X}_a \in \mathbb{R}^{128\times T\times F}$.
- *Visual front-end.* A single-channel clip $\mathbf{V} \in \mathbb{R}^{1\times T_v\times 96\times 96}$ passes through:
  1. 3-D Conv (5,7,7) (1→64) with stride $(1, 2, 2)$;
  2. MaxPool (1,3,3) $\Rightarrow T_v = T$, $H = W = 12$;
  3. three ResNet layers (64-64-128) acting per frame;
  4. a five-layer temporal CNN built from depth-separable 1-D blocks.

  This yields $\mathbf{Z} \in \mathbb{R}^{64\times T}$.
- *FiLM conditioning.* A 1×1 Conv1D doubles $\mathbf{Z}$ to $(\gamma_{\text{raw}}, \beta)$. After rescaling $\gamma = 1 + \gamma_{\text{raw}}$ and broadcasting along F, feature-wise modulation is applied: $\tilde{\mathbf{X}}_a = \gamma \odot \mathbf{X}_a + \beta$.
- *Refinement.* A 1×1 Conv2D (128→128) + PReLU completes the stem, producing $\mathbf{H}^{(0)} \in \mathbb{R}^{128\times T\times F}$.

Local–Global Transformer stack. Four identical *Loco-Transformer* blocks propagate $\mathbf{H}^{(0)} \mapsto \mathbf{H}^{(4)}$. Each block contains:

1. **Intra-frequency module**: Conv-FFN (kernel 4, hidden 192) → $\text{MHSA}_{\text{freq}}$ (8 heads, 128 dim) with rotary positional embedding; RMS-GroupNorm and residual scaling (0.5).
2. **Inter-time module**: identical Conv-FFN + $\text{MHSA}_{\text{time}}$.
3. **Frame aggregator**: mean over F → LayerNorm → temporal MHSA (4 heads).

RMS normalisation prevents FP16 overflow; sequence lengths up to $T = 750$ run without gradient scale drops.

Output heads. Three deconvolutions (kernel 3, pad 1) act on $\mathbf{H}^{(4)}$:

- complex mask $\hat{\mathbf{M}}_c \in \mathbb{R}^{2\times T\times F}$,
- magnitude mask $\hat{\mathbf{M}}_m = \sigma_{\beta=1.2}(\cdot) \in [0, 1.2]^{1\times T\times F}$,
- phase unit vector $(\sin\hat{\varphi}, \cos\hat{\varphi}) \in \mathbb{R}^{2\times T\times F}$.

Enhanced STFT: $\hat{\mathbf{S}} = \hat{\mathbf{M}}_c \odot \mathbf{X} + \hat{\mathbf{M}}_m \odot |\mathbf{X}|e^{j\hat{\varphi}}$, followed by inverse STFT.

Complexity. Total parameters: ≈ 6.1 M.

2 Training Procedure

Twenty epochs on the official training split with AdamW (lr 3×10^{-4}, 5 k warm-up, weight-decay 10^{-2}), automatic mixed precision and gradient clipping (5). Objective = $\text{SI-SDR}_{\text{neg}} + \lambda = 0.5\cdot L_1$ spectral loss. Training uses two 2080TI GPUs, batch 4, and reaches 1.2× real-time per GPU.

3 Development-Set Results

System	SI-SDR↑	PESQ↑	STOI↑
AV-LocoFiLM	**10.9 dB**	**2.20**	**0.84**