Trend Lab @ University of Stirling

RL-Optimized TCN for Efficient Audio-Visual Speech Enhancement: A NAS-Driven Approach

1. System Overview

Our audio-visual speech enhancement system combines visual lip movements with noisy speech signals to reconstruct clean audio. The key innovation is a neural architecture search (NAS) optimized model that achieves **10× parameter reduction** (22M → 2.2M) while maintaining near-baseline performance. The system can operate in real-time on edge devices.

2. Neural Architecture Search with Reinforcement Learning

We employed reinforcement learning to automatically discover efficient architectures:

Search Space:

- Encoder channels: {128, 192, 256}
- Bottleneck dimension: {128, 192, 256}
- TCN repeats: {2, 3, 4}
- TCN blocks per repeat: {4, 6, 8}
- Visual feature dimension: {128, 256}

Agent Design:

- Policy network: 2-layer LSTM
- Action: Modify one architectural dimension
- Reward function:
  Reward = SI-SNR improvement - 0.7 × Parameter overuse penalty

Search Process:

1. Start from baseline 22M-parameter model
2. Agent proposes architecture modifications
3. Train candidate for 10 epochs (accelerated evaluation)
4. Compute reward based on performance/size tradeoff
5. Update agent using Proximal Policy Optimization (PPO)
6. Repeat for 35 cycles (total search cost: 350 GPU-hours)

3. Final Model Architecture (2.2M Parameters)

3.1 Visual Pathway (0.3M params)

- Input: 112×112 lip ROI at 25fps
- 3D convolution (5×5×5 kernel) + max pooling
- Depthwise separable 2D convolutions
- Output: 128-dimensional temporal features

3.2 Audio Encoder

- Input: 16kHz audio (3-second segments)
- 1D convolution: Kernel=40ms, Stride=20ms
- Output channels: 192
- Output representation: Time-frequency embedding

3.3 NAS-Optimized Separator (1.5M params)

- **Visual projection**: 128 → 192 channels
- **Temporal processing**:
    - 3 repeats of 6 temporal blocks
    - Dilations: 1, 2, 4, 8, 16, 32 per repeat
    - Hidden size: 384 (2× expansion ratio)
- **Feature fusion**: Audio + visual features via addition
- **Mask generation**: Adaptive masking layer

3.4 Audio Decoder

- Basis signal reconstruction via linear layer
- Overlap-add synthesis (20ms frame step)
- Output: Enhanced 16kHz waveform

4. Key Architectural Insights from NAS

1. **Visual compression**: Features reducible to 128 dims (50% savings)
2. **TCN efficiency**: Optimal at 3×6 blocks (43% reduction)
3. **Bottleneck dimension**: 192 channels balances information flow
4. **Expansion ratio**: 2.0 provides optimal compute/accuracy tradeoff

5. Training Configuration

- **Optimization**: Adam (lr=1e-3)
- **Scheduling**: Reduce-on-plateau (factor=0.8, patience=3)
- **Loss function**: Scale-invariant SNR maximization

Advantages

1. **Edge deployable**: 8.8MB model size fits mobile constraints
2. **Minimal quality drop**: <1dB degradation from 22M baseline
3. **Efficient search**: 92% more efficient than random architecture exploration

8. Conclusion

Our NAS-optimized AVSE model demonstrates that reinforcement learning can effectively discover efficient architectures for multimodal speech enhancement. The 2.2M-parameter solution maintains robust performance while enabling real-time operation on resource-constrained devices, making it ideal for hearing aids, video conferencing, and augmented reality applications. The NAS framework achieved 10× model compression with only 35 architecture evaluations, providing a blueprint for efficient neural design in multimodal systems.