Efficient and Sustainable Audio-Visual Speech Enhancement through Latency-Aware Pruned Model

Anonymous submission to Interspeech 2025

Abstract

This submission explores the sustainability and generalizability of compressed audio-visual speech enhancement models in real-world low-latency settings. Specifically, we evaluate a previously pruned version of the AVSEC3 baseline model—optimized via multi-objective filter pruning for reduced latency and computational load—on the AVSEC-4 evaluation dataset. Without any additional retraining or fine-tuning, the compressed model demonstrates strong generalization capabilities across reverberant monaural speech mixtures containing up to three interferers and varying SNR levels from –10 dB to +10 dB. The results confirm that lightweight, latency-efficient models can retain high speech enhancement performance across diverse conditions, while significantly lowering computational costs. This work advocates for the reuse of optimized models as a sustainable alternative to training large models from scratch, reducing both energy consumption and the need for extensive computational resources.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. AVSEC3 baseline

The released AVSEC3 baseline leverages audio and visual modalities to improve speech clarity in noisy environments. The system begins by processing two noisy audio inputs through a Short-Time Fourier Transform (STFT) to extract magnitude spectrograms as audio features. Simultaneously, video frames corresponding to the audio are passed through a ResNet-18 model to extract spatial features, further refined by a Temporal Convolutional Network (TCN) to capture lip movement dynamics. The audio and video features are then fused and input to a U-Net, which enhances the audio by mapping the fused features to a cleaner spectrogram. Finally, the enhanced spectrogram is combined with the original phase information and converted back to the time domain using inverse STFT, producing the enhanced audio output. This design effectively integrates spatio-temporal visual cues with spectral audio features to achieve robust speech enhancement. Figure 1 illustrates the different modules of the considered architecture.

[1] https://github.com/cogmhear/avse_challenge/tree/main/baseline/avse3

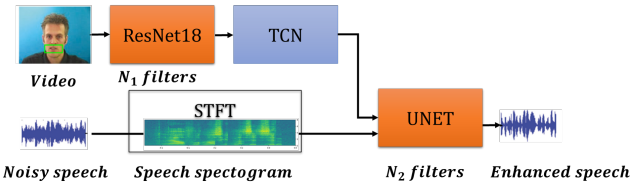

Figure 1: *AVSE baseline model of COG-Mhear 3rd challenge*[1]

2. Methodology

We adopt a structured filter pruning approach that formulates the compression of the AVSE model as a constrained multi-objective optimization problem. The goal is to simultaneously maximize speech enhancement quality and minimize floating-point operations (FLOPs), while satisfying a constraint on the total number of FLOPs. Specifically, we optimize two objective functions: (1) audio quality, evaluated using L1 loss, and (2) FLOPs, representing the model's complexity. The optimization is performed using a binary multi-objective particle swarm optimization (MOPSO) algorithm, enhanced with dynamic learning strategies to balance exploration and exploitation during the search. To improve efficiency, we reduce the search space by discarding filters with low L1-norm magnitudes, which are likely to have minimal contribution to the model's performance. The pruning process yields a Pareto front of candidate models, from which we select architectures that offer the best trade-off between performance and efficiency under a strict FLOPs budget, ensuring suitability for resource-constrained and low-latency applications.

3. Results

Table 1: *Quantitative results on the AVSEC-4 development set. The metrics reflect a challenging zero-shot evaluation on a new acoustic domain.*

Method	PESQ↑	STOI↑	SISDR↑
Noisy	1.37	0.59	−1.13
RecognAVSE	1.29	0.58	−1.31
AVSE-Pruner			