MSF-AVSE: Multi-Stream Fusion Audio-Visual Speech Enhancement

The proposed **MSF-AVSE** (Multi-Stream Fusion Audio-Visual Speech Enhancement) model is a novel architecture designed to leverage both spatial audio cues from a two-microphone setup and visual information from lip movements to enhance speech in noisy environments. The model is composed of four core modules: a *Spatial Audio Encoder*, a *Visual Encoder*, a *Cross-Modal Fusion Block*, and a *Mask Estimator and Decoder*. The Spatial Audio Encoder extracts magnitude and phase-related spatial features, including Interchannel Phase Difference (IPD), from the stereo audio input and processes them through convolutional layers to form a compact audio representation. The Visual Encoder captures dynamic lip movement patterns using a 3D convolutional neural network followed by a temporal aggregation via Bi-directional LSTM, aligning visual information temporally with audio. These two modalities are then integrated in the Cross-Modal Fusion Block using a self-attention-based fusion mechanism that enables adaptive weighting of features from both domains. Finally, the fused representation is passed to a Mask Estimator that predicts a complex ratio mask, which is applied to the noisy spectrogram to reconstruct the enhanced speech using an inverse Short-Time Fourier Transform (iSTFT). This architecture effectively exploits spatial, spectral, and visual cues, resulting in improved speech enhancement performance under challenging noise conditions.