A COMPLEX IDEAL RATIO MASK BASED AUDIO-VISUAL SPEECH ENHANCEMENT FOR THE 2ND COG-MHEAR AVSE CHALLENGE

1st Feixiang Wang
Institute of Computing Technology, CAS
University of Chinese Academy of Sciences (UCAS)
Beijing, China

2nd Shuang Yang
Institute of Computing Technology, CAS
Beijing, China

We propose a complex spectral ideal ratio mask based audio-visual speech enhancement (AVSE) model for the 2nd COG-MHEAR AVSE challenge. The model employ the noisy audio, target speakers face to estimate a complex ideal ratio mask (cIRM).

Our model consists of a visual branch and a U-Net style audio branch.

I. Visual Brach

Our visual branch is built upon a 3D convolutional layer, which downsamples the input image sequence in the spatial domain. To optimize the model convergence without sacrificing performance, we incorporate a lightweight ShuffleNet V2 network. We added a Spatial Attention Module in the middle of ShuffleNet V2, which is proven to effectively capture global facial features. Additionally, a Temporal Convolutional Network (TCN) is utilized to capture temporal dependencies from the output features of ShuffleNet V2. The resulting visual features from TCN have a dimension of $C_v \times T_v$, where C_v represents the channel dimension, T_v denotes the temporal dimension.

II. Audio Branch

For the audio branch, we employ a U-Net style network.

A. U-Net Encoder

The U-Net encoder takes the complex spectrum S_{noisy} derived from the Short-Time Fourier Transform (STFT) of the noisy audio s_{noisy} as its input. S_{noisy} has dimensions of $2 \times F \times T$, with F and T representing the frequency and time dimensions of the spectrum, respectively. The encoder consists of 9 layers of convolutional networks and average pooling layers, which downsamples the input spectrum's frequency dimension to 1 and the time dimension to T_a. The resulting output feature has a dimension of $C_a \times T_a$, where C_a represents the channel dimension and T_a represents the time dimension.

B. Bottleneck

The output features of the visual branch and audio branch are fused through a learned weight vector using weighted addition.

C. U-Net Decoder

The decoder employs a symmetric structure to the encoder. It takes the fused audio-visual features as input and undergoes a series of upsampling operations to generate a predicted cIRM M_p with dimensions of $2 \times F \times T$, matching the input spectrum's dimensions. The predicted M_p is multiplied with the input spectrogram in the complex domain to obtain the predicted complex spectrogram. Finally, the enhanced audio is obtained by performing the inverse Short-Time Fourier Transform (iSTFT) on the predicted complex spectrogram.

III. Experiment

A. Training Details

We conducted evaluation on the provided dataset and employed a three-stage training approach for a total of 490 epochs (400 + 60 + 30). In the first stage, we performed overall training. During the second stage, we froze the U-Net Encoder and partially fine-tuned the remaining Visual Branch. In the third stage, we unfroze the entire model and fine-tuned it. The model has a total parameter count of 59.02M. We utilized the One-Cycle Cosine learning rate policy and applied linear warm-up for the first 10% of epochs in each stage. The initial learning rate is set 3e-4.

B. Results

The test results on the evaluation set are shown in the table I

Metrics	PESQ	STOI	SI-SDR
noisy	1.160	0.638	-4.696
speech noise	1.737	0.834	6.678
non-speech noise	1.811	0.830	8.618
average	1.775	0.832	7.655

TABLE I
COMPARISON WITH AV C-REF ON GRID DATASET.