AUDIO-VISUAL SPEECH ENHANCEMENT METHOD BASED ON CONTINUOUS FUSION OF SPEECH FEATURES FOR 2ND COG-MHEAR AVSE CHALLENGE

Wen-Ze Ren[*] , *Yu Tsao*[†]

[*]National Taiwan University , [†]Academia Sinica

1. PROPOSED APPROACH

Our proposed method is an audio-visual speech enhancement neural network based on layer-by-layer fusion of speech features as shown in Fig.1. The entry associated with this report is **rezzsl_new_2**. First, the visual signal is processed by Resnet. The speech features will be extracted by 1D-Resnet, and we retain the speech features x_0, x_1, x_2, and x_3 of different dimensions in each layer. After the audio and visual features are fused, a deep LSTM neural network is used to train a mask for speech enhancement, and the original speech signal is multiplied by the mask to obtain the input of the Audio Decoder. Compared with the traditional voice-enhanced Decoder that restores the voice signal, we have made an innovation here. In order to make better use of the feature information of each dimension of the embeddings on the Audio Encoder, we consider merging with the input of the Audio Decoder. The x_3, x_2, x_1, and x_0 returned by the Audio Encoder will be fused with the input of the Audio Decoder layer by layer. After the fusion, an LSTM neural network with a small number of layers will be trained separately to perform noise reduction again, and finally, a clean audio file will be obtained after deconvolution.

2. EXPERIMENTAL SETUP AND RESULTS

2.1. Experimental Setup

The training data set used by the model is the LRS3 data set officially provided by AVSE Challenge. The total trainable parameters of the model are 19.9M. Batch size is set to 4, learning rate is 0.001, a total of 15 epochs are trained. The training time of each epoch is about 40 minutes, and the total training time is 9 hours. If the performance of the model does not improve within two epochs, we apply a learning rate decay by a factor of 0.8. The training was conducted on a machine equipped with an RTX3080 graphics card with 24Gb of video memory. The video memory occupied during the training process remained at about 21Gb.

2.2. Experimental Results

Our findings indicate a considerable improvement in PESQ, STOI, and SI-SDR scores compared to those derived from the noise-mixed dataset, despite the model being trained for only 15 epochs. The most pronounced enhancements are observed in the STOI and SI-SDR scores, with the STOI improving by 25%, and the SI-SDR demonstrating an 8dB increase compared to the original values.

It is shown that this layer-by-layer feature fusion fully utilizes the good performance of multi-dimensional feature information, which suggests to see the potential of this continuous fusion of speech features.

	PESQ	STOI	SI-SDR
noisy	1.13	0.44	-5.07
ours	**1.26**	**0.50**	**3.00**

Table 1. Compare scores on the evaluation dataset

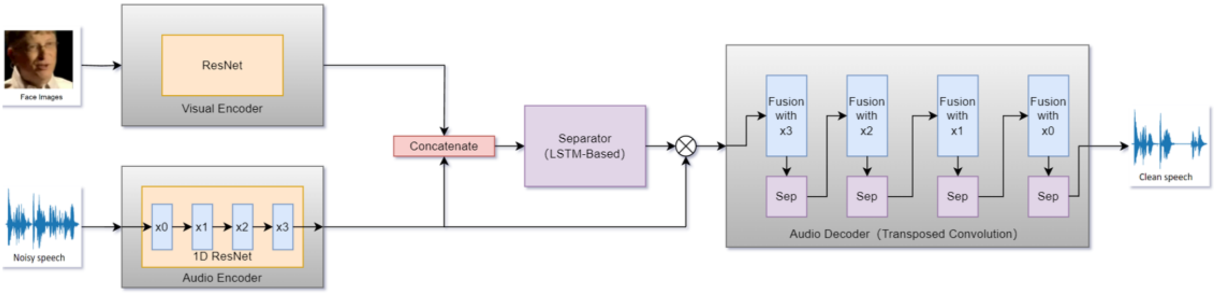

Fig. 1. Architecture of our AVSE Model