DEEP COMPLEX U-NET WITH CONFORMER FOR AUDIO-VISUAL SPEECH ENHANCEMENT

Shafique Ahmed[1,2], *Chia-Wei Chen*[3], *Wenze Ren*[3], *Chin-Jou Li*[3], *Ernie Chu*[3], *Jun-Cheng Chen*[1], *Amir Hussain*[4], *Hsin-Min Wang*[1], *Yu Tsao*[1], *and Jen-Cheng Hou*[1]

[1]Academia Sinica, Taiwan [2]National Tsing Hua University, Taiwan [3]National Taiwan University, Taiwan [4]Edinburgh Napier University, UK

1. INTRODUCTION

For the challenge, we propose a novel approach for audio-visual speech enhancement (AVSE) that combines the deep complex U-Net architecture and Conformer blocks. Our AVSE framework utilizes lip movement cues as visual features to enhance the AVSE process. resulting in improved speech intelligibility and quality.
The entry associated with this report is BioASP_CITI.

As depicted in Figure 1, our AVSE model integrates lip movement cues with audio features by employing a ResNet-18 architecture to extract relevant information from lip movement videos. These visual features are then concatenated with the audio features obtained from a complex encoder. To ensure effective integration and fusion of the two modalities, we utilize a temporal upsampling technique, aligning the temporal resolution of visual features with that of the audio features.

To capture long-range dependencies and contextual information, we replace LSTM layers in the deep complex U-Net architecture with Conformer blocks. This enhancement allows the model to effectively capture fine-grained information from both audio and visual modalities, leading to superior AVSE performance. Our proposed approach showcases the potential of leveraging audio-visual cues for enhancing speech quality and intelligibility.

2. EXPERIMENTAL SETUP AND RESULTS

In our AVSE model, we used Convolutional Short-Time Fourier Transform (ConvSTFT) for audio preprocessing and ResNet-18 for visual feature extraction with temporal upsampling. Additionally, the model incorporated 5 complex Conv2D blocks for audio processing, 2 Conformer blocks for capturing audio-visual dependencies, and decoder blocks to reconstruct the complex mask. Training was conducted using Scale-Invariant Signal-to-Noise Ratio (SI-SNR) as the loss function.

The evaluation compared three speech types: noisy speech, enhanced speech from baseline models, and enhanced speech from our proposed models. We measured the performance using standard metrics, Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI).

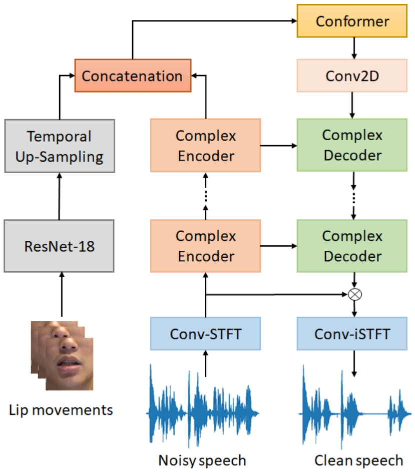

Fig. 1. Architecture of our AVSE model.

The experimental results, as shown in Table 1, demonstrated that the enhanced speech from all AVSE models surpassed noisy speech in terms of both quality and intelligibility. Our Transformer- and Conformer-based models outperformed the baseline approach. Notably, the Conformer-based model achieved the highest performance across both evaluation metrics.

System Description: Our AVSE model is equipped with 20M trainable parameters and was trained on a Tesla V100 GPU. The training process spanned 120 epochs and took approximately 48 hours to complete. For inference, the model's CPU computation averaged around 60ms, while utilizing the GPU reduced the average processing time to 50ms.

Table 1. Objective assessment scores of the noisy speech and the enhanced speech of the baseline and the proposed models.

	PESQ	STOI
Noisy	1.15	0.64
Baseline	1.70	0.83
Ours (Transformer)	1.79	0.83
Ours (Conformer)	**1.84**	**0.84**