A DCCRN-BASED AUDIO-VISUAL SPEECH ENHANCEMENT APPROACH FOR THE 1ST COG-MHEAR AVSE CHALLENGE

Jen-Cheng Hou[1], *Ryandhimas Zezario*[2], *Chin-Jou Li*[2], *I-Chun Chern*[3], *Jun-Cheng Chen*[1], *and Yu Tsao*[1]

[1]Academia Sinica, Taiwan [2]National Taiwan University, Taiwan [3]Carnegie Mellon University, USA

1. PROPOSED APPROACH

Our proposed approach is based on deep complex convolutoin recurrent network (DCCRN) [1], which has been shown to be effective for speech enhancement (SE) by predicting target complex spectrum via complex-valued operation. We incorporate additional visual information (e.g. lip motion) into a DCCRN model as an audio-visual speech enhancement (AVSE) approach for the 1st COG-MHEAR AVSE Challenge. The entry associated with this report is `BioASP_CITI`.

Our proposed model, AV-DCCRN, is shown in Figure 1. Noisy speech is converted into complex spectrum and processed via a U-net like complex encoder-decoder architecture, where the latent representations are processed with a bi-directional long short-term network (BLSTM). The visual features of the lip motion and the speech features are fed into cross-attention modules, which consist of multi-head attention (MHA), to better couple the audio-visual features. The training objective is to minimize a weighted sum of the scale-invariant signal-to-noise ratio (SI-SNR) loss and the $L1$ loss between the predicted speech and the clean speech.

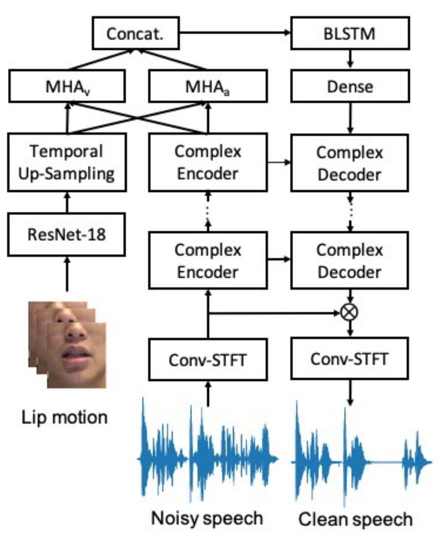

Fig. 1. Our proposed AV-DCCRN model. MHA represents Multi-Head Attention.

2. EXPERIMENTAL RESULTS

The experimental results on the development set is shown in Table 1. We can observe that the inclusion of MHA blocks for audio-visual feature coupling can obtain a performance boost compared to naive cross-modal feature fusion. Our experimental results show the proposed model can significantly outperform the baseline model with a 40.7% and 23.8% improvement in perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI), respectively. The number of trainable parameters of the proposed model is 18.6M.

Methods	PESQ	STOI
Noisy speech	1.15	0.64
Baseline	1.30	0.67
AV-DCCRN (w/o MHA)	1.62	0.81
AV-DCCRN (w/ MHA)	**1.80**	**0.83**

Table 1. Comparison of the evaluation scores of the enhanced speech on the development set. The baseline model is the one provided by the challenge organizers.

3. REFERENCES

[1] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *INTERSPEECH*, 2020.