CHINESE ACADEMY OF SCIENCES
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

BMVC 2023

Cooperative Dual Attention for Audio-Visual Speech Enhancement with Facial Cues

Feixiang Wang, Shuang Yang, Shiguang Shan, Xilin Chen

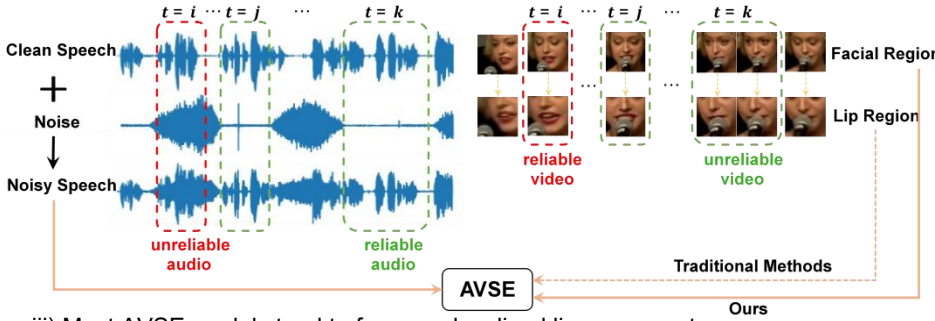
Motivation

- Most existing audio-visual speech enhancement (AVSE) methods utilize lip movements to assist in enhancing target speech. The facial region beyond lip contains abundant speech-related information, such as gender, nationality, skin color, and so on, which can reflect the speaker's timbre and accent.

Challenges

i) The facial region contains non-speech-related information.

ii) Both visual and audio quality fluctuate in real-world speech scenarios.



iii) Most AVSE models tend to focus on localized lip movements.

Experimental Data

- LRS3 (Audio-Visual Speech Dataset)

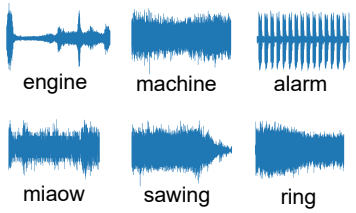
Pose Variation

Lip Occlusion

⋮

Low Resolution

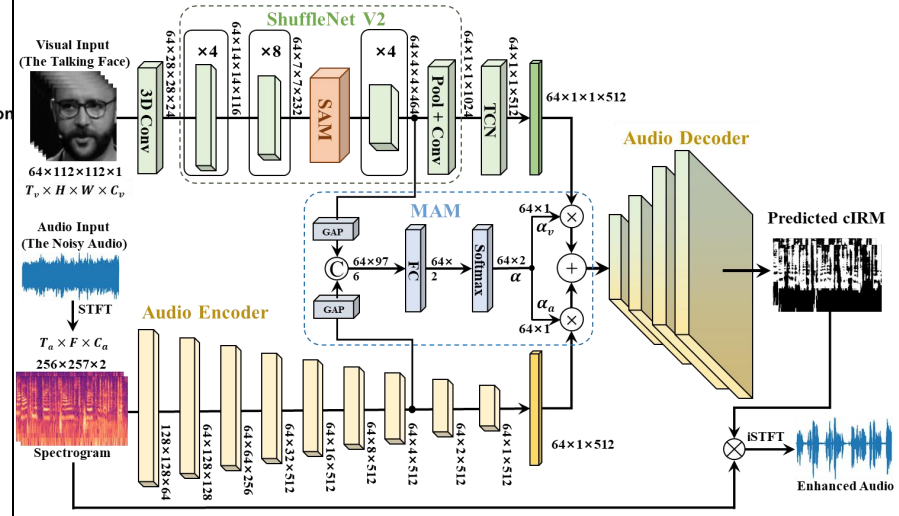
- DNS4 (Audio Noise)



...

The Proposed DualAVSE

- **Spatial attention module (SAM)** captures global facial information with a self-attention layer on the feature map.
- **Modality attention module (MAM)** dynamically perform modality fusion with an attention vector to measure the reliability of audio and visual modalities.



Ablation Study

- **AOSE vs. AVSE:** Visual modality improves performance.
- **MAM & SAM:** Both MAM and SAM are effective.

Input \ Metrics	SDR	PESQ	STOI
AOSE Baseline	11.09	1.999	0.868
AVSE Baseline	11.41	2.071	0.873
+MAM	11.70	2.103	0.879
+SAM	12.12	2.186	0.887
DualAVSE	**12.32**	**2.241**	**0.889**

Experiments on LRS3 + DNS4 datasets

Face vs. Lip

- Face input outperforms lip and is more robust.

Input \ Metrics	SDR	PESQ	STOI
Reliable Face	**12.32**	**2.241**	**0.889**
Mask Whole Face	12.09	2.220	0.886
Mask Lip in Face	12.23	2.223	0.888
Randomly Mask Face	12.27	2.236	0.889
Reliable Lip	12.11	2.190	0.888
Mask Whole Lip	11.90	2.169	0.884
Randomly Mask Lip	12.08	2.186	0.887

Comparison with others

Model \ Metrics	SDR	PESQ	STOI
DEMUCS [1]	11.85	1.631	0.839
MuSE [2]	8.53	1.460	0.797
VisualVoice [3]	10.32	1.963	0.865
DualAVSE	**12.32**	**2.241**	**0.889**

Comparison on LRS3 + DNS4 datasets

Qualitative Results

- SAM enables model to capture global information from face.

Acknowledgement: This work is partially supported by National Natural Science Foundation of China (No. 62276247, 62076250).

Reference:

[1] Défossez, A., Synnaeve, G., Adi, Y. (2020) Real Time Speech Enhancement in the Waveform Domain. Proc. Interspeech 2020, 3291-3295, doi: 10.21437/Interspeech.2020-2409.

[2] Pan Z, Tao R, Xu C, et al. Muse: Multi-modal target speaker extraction with visual cues[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6678-6682.

[3] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15490–15500. IEEE, 2021.