

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

Feixiang Wang^{1,2}

wangfeixiang19@mailsucas.ac.cn

Shuang Yang^{1,2}

shuang.yang@ict.ac.cn

Shiguang Shan^{1,2}

sgshan@ict.ac.cn

Xilin Chen^{1,2}

xlchen@ict.ac.cn

¹ Key Laboratory of Intelligent

Information Processing of Chinese

Academy of Sciences (CAS),

Institute of Computing Technology,

CAS,

Beijing 100190, China

² University of Chinese Academy of

Sciences,

Beijing 100049, China

In this supplementary material, we provide further quantitative and qualitative analysis, together with more comparison results with other methods. In Sec A, we present several randomly generated examples to show the effect of our model. Sec B presents a visual analysis of our ablation study results. Sec C and Sec D provide a further qualitative analysis of our proposed DualAVSE model. In Sec E, we perform a detailed comparison with several state-of-the-art (SOTA) audio-visual speech enhancement (AVSE) methods on 3 benchmarks.

A Audio-Visual Examples

To clearly demonstrate the effectiveness of our DualAVSE model in speech enhancement, we have created a video. This video features both the original audio-visual data containing noisy speech and the enhanced output produced by our model.

B Further Analysis of Our Ablation Study Results

Besides the tabulated ablation results in Table 1 in the main text, we plot the results of the ablation study into a histogram here. Unless explicitly indicated as "lip," the AVSE models are based on the face region by default. As shown in Figure 1, the results reveal several key observations.

Firstly, introducing visual modality information, whether from the lip region or the face region, significantly improves the performance of speech enhancement compared with the audio-only speech enhancement (AOSE) Baseline. Particularly, the models that utilize the face region as input consistently outperform the models that utilize the lip region under the same conditions. This may seem contradictory to the fact mentioned in the submission paper

that not exactly utilizing facial information could potentially result in performance degradation. However, it is important to note that the baseline model used in our study incorporates a state-of-the-art lip-reading frontend. It can accurately extract semantic information from the lip movements in face videos, which is illustrated in Sec C.

Secondly, both the modality attention module (MAM) and spatial attention module (SAM) contribute to the improvement of the AVSE task. The SAM module leads to even more significant performance improvements due to its capability to capture more visual speech information based on facial cues.

Lastly, when incorporating both SAM and MAM modules, our proposed DualAVSE demonstrates more significant performance improvements, as measured by PESQ and STOI metrics. The results show that the integration of these modules enables more effective utilization of both visual and audio modalities, leading to improved speech quality. Specifically, the DualAVSE model addresses two key aspects: first, it captures rich and valuable information beyond the lip region by considering the entire face; second, it addresses the challenge of the varying reliability of audio-visual modality information in real-world scenarios over time.

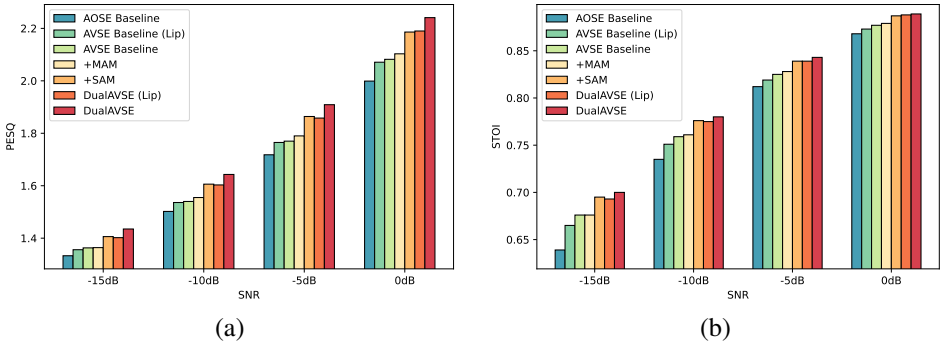


Figure 1: Bar plots of the ablation study results on PESQ and STOI metrics. The ablation study involves the evaluation of submodules and model inputs. (a) PESQ results. (b) STOI results.

C Visualization of the Capability to Learn Facial Cues of DualAVSE Model

We employ the Grad-CAM [14] method to visualize the heat maps of the intermediate features in the visual branch. Specifically, we focus on the output features of the third module of the ShuffleNet V2 in the visual branch. These features are located in the later layers of the visual branch, where they have already captured rich spatial and semantic information. We performed the visualization analysis separately for the AVSE baseline model with the face region as input (BASE) and the model with the SAM module incorporated (+SAM).

As depicted in Figure 2, the BASE model shows a pronounced focus on the region surrounding the speaker’s lips, consistent with findings in previous studies [20, 21]. This finding suggests that the BASE model effectively captures and utilizes lip motion information for speech enhancement. As shown in Figure C, the +SAM model exhibits a tendency to extract valuable global information from the entire face region rather than solely focusing on localized lip movements. This observation highlights the capability of the +SAM model to effectively utilize facial information, resulting in enhanced performance in AVSE.

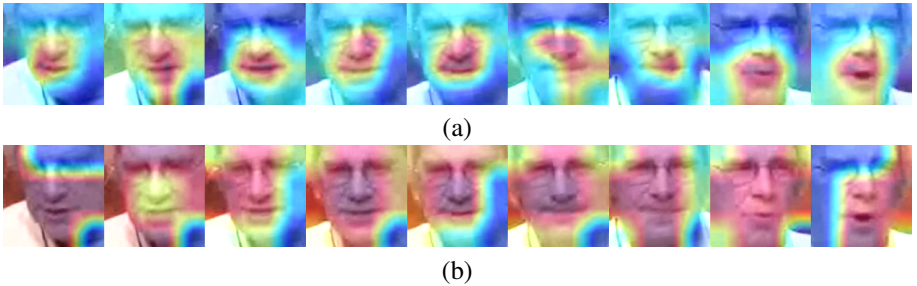


Figure 2: **Visualization analysis of intermediate layer features in the visual branch, using a color scheme ranging from blue (indicating low contribution) to red (indicating high contribution).** (a) Heat maps of the BASE model. (b) Heat maps of the +SAM model.

D Analysis of Unreliable Modalities

We corrupt the inputs with random masks to simulate unreliable modality and then visualize the bottleneck-layer features f_{av} . The bottleneck-layer features are averaged along the channel dimension, resulting in a one-dimensional vector of the same length as the video T_v . As shown in Figure 3, when one modality is unreliable, the fused features become closer to the features of the other and indeed rely more on the other reliable modality. To elaborate, consider Figure 3. (a) and (b), when we masked a specific segment of one modality’s input, i.e. the first 30% of the visual feature in (a), the rising and falling trends in the fused bottleneck-layer features closely resembled the features of the unmasked modality, i.e. the audio feature. Moreover, when one modality was entirely masked as shown in Figure 3. (c) and (d), i.e. the visual modality in (d), the bottleneck-layer features only exhibited changes in accordance with the features of the unmasked modality, i.e. the audio modality.

E More Detailed Comparisons with the SOTA Methods

In addition to the comparisons presented in the main text, We perform a more detailed comparison with the SOTA methods on 3 benchmark datasets in this section. It is important to note that the specific methods compared may differ for each dataset due to variations in the SOTA methods across different datasets. This allows for a thorough examination of our approach’s performance in different scenarios and ensures a fair comparison with the best-performing methods on each dataset.

GRID [3], CHiME3 [2]: We compare the proposed DualAVSE model with the SOTA AVSE approaches on GRID datasets. Following [17], we utilize the noises from the CHiME3 dataset to synthesize the noisy input audios and perform an evaluation with the test signal-to-noise ratio (SNR) levels of both -5dB and 0dB. As shown in Table 1, the DualAVSE model achieves the best performance in both the PESQ improvement (PESQi) and STOI improvement (STOIi) metrics with different test SNR levels.

TCD-TIMIT [7], NTCD-TIMIT [1]: We further evaluated our DualAVSE model on the TCD-TIMIT dataset, comparing it with SOTA AVSE methods. The TCD-TIMIT dataset consists of AV speech data from 56 English speakers with an Irish accent. Each utterance is approximately 5 seconds long and sampled at 16kHz. As recommended in [7], we split the dataset into training, validation, and testing sets, with 39 speakers for training, 8 for validation, and 9 for testing. The noisy speech input is derived from the NTCD-TIMIT

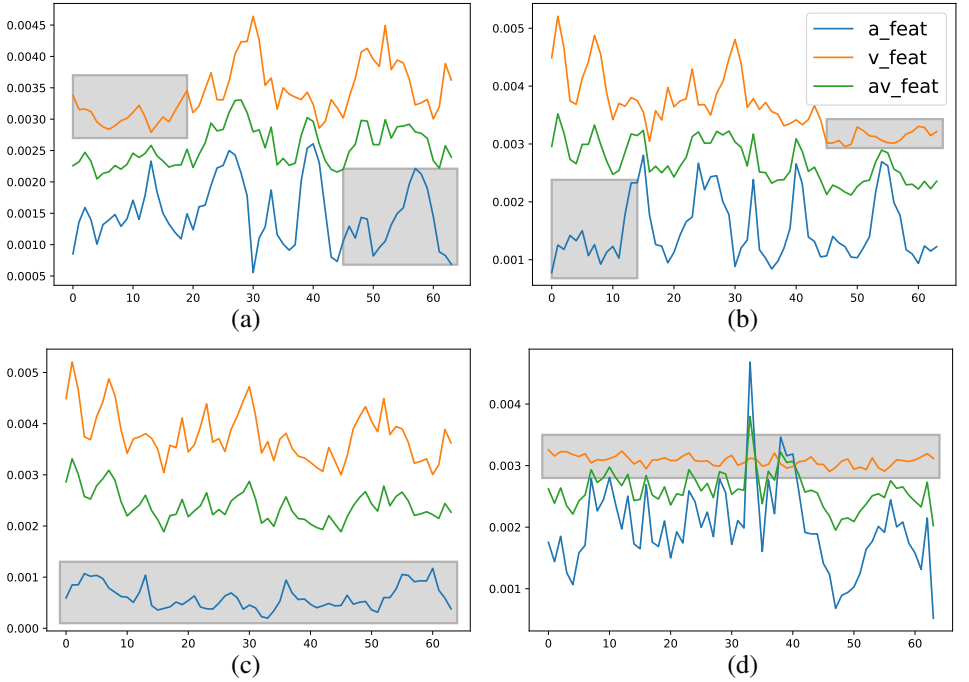


Figure 3: Visualization of the bottleneck features. The signals with gray boxes are masked. (a) Mask the first 30% of the audio modality and the last 30% of the visual modality; (b) Mask the last 30% of the audio modality and the first 30% of the visual modality; (c) Mask the whole audio modality; (d) Mask the whole visual modality.

dataset. This dataset is created by adding six different types of noise to the original speech data from the TCD-TIMIT corpus. The noise types include Living Room, White, Cafe, Car, Bable, and Street, and each noise type is associated with a specific SNR. Similar to the approach in [6], we selected 5 utterances per noise level and noise type for each test speaker to create a test set of 1350 utterances. The reporting metrics in [6] contains SI-SDR [9], PESQ, and STOI. We report the score improvement as a means of comparison. As shown in Table 2, the proposed DualAVSE model achieves the best performance across all metrics at most SNR levels, except for the STOI metric at SNR 15dB. This phenomenon is mainly caused by the fact that our input noisy speech already has a higher STOI score of 0.80 at SNR 15dB, while theirs is 0.69. This significant difference in the initial STOI scores makes it more challenging for our model to further improve the STOI score compared with the other methods.

MEAD [16], DEMAND [15]: We conducted a comparison between our DualAVSE model and the SOTA method [8] on the MEAD dataset. This dataset consists of recordings from 46 participants, who uttered sentences expressing eight different emotions at three intensity levels under seven camera viewpoints. To ensure a fair comparison, we followed the same selection criteria as [8], choosing videos that captured frontal views and the highest level (level 3) of emotion intensity. These selected videos often exhibit significant head movements and exaggerated lip motions, posing challenges for AVSE. For training, we utilized approximately 5 hours of videos from the MEAD dataset. Additionally, 0.7 hours were

Metrics	STOIi(%)		PESQi	
	-5	0	-5	0
SNR(dB)				
L2L [4]	11.14	8.86	0.54	0.62
VSE [5]	-	-	0.45	0.60
OVA [17]	-	-	0.40	0.66
VSET [11]	-	-	0.50	0.75
MHCA-AVCRN [19]	13.51	11.25	0.76	0.88
M3Net [18]	13.42	11.31	0.75	0.89
DualAVSE	15.79	13.56	0.76	0.92

Table 1: Results on GRID dataset.

Metrics	SI-SDRi(dB)					PESQi					STOIi				
	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
A-VAE [12]	4.34	5.12	5.93	6.07	5.76	0.16	0.19	0.20	0.21	0.05	0.02	0.02	0.04	0.04	0.04
AV-VAE [12]	6.15	6.86	7.38	7.22	6.52	0.24	0.27	0.29	0.28	0.08	0.02	0.03	0.04	0.05	0.04
A-DKF [6]	5.78	6.80	7.67	8.35	7.71	0.27	0.32	0.36	0.38	0.18	0.02	0.05	0.07	0.09	0.08
AV-DKF [6]	9.02	9.50	10.10	9.62	8.56	0.43	0.48	0.49	0.43	0.20	0.05	0.08	0.09	0.10	0.08
DualAVSE	18.50	17.18	15.35	12.93	10.71	0.45	0.67	0.88	1.06	1.16	0.15	0.15	0.13	0.10	0.06

Table 2: Results on TCD-TIMIT. Higher is better for all metrics.

reserved for validation, and another 0.7 hours were allocated for testing purposes. Following [8], we utilize noise from the DEMAND dataset to synthesize the mixture input audios. Similarly to our experiments on the TCD-TIMIT dataset, we computed the improvement in SI-SDR, PESQ, and STOI metrics to ensure a fair comparison. The results presented in Table 3 demonstrate that our proposed DualAVSE model outperforms all other methods in terms of all evaluated metrics across various SNR conditions. It is worth mentioning that the competing models fail to show any improvement in STOI at the SNR of 10dB, whereas DualAVSE still manages to enhance the STOI even at such a high SNR level. For example, the Res-AV-CVAE-RFF model [8] shows an STOIi of -0.01 at the SNR of 10dB, whereas our DualAVSE model achieves an STOIi of 0.03.

Metrics	SI-SDRi(dB)					PESQi					STOIi				
	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10
A-VAE [10]	8.91	10.33	10.52	9.81	8.14	0.03	0.27	0.35	0.38	0.31	0.01	0.03	0.04	0.01	-0.01
AV-CVAE [13]	8.96	10.58	10.45	9.46	7.65	0.12	0.32	0.39	0.37	0.31	0.02	0.04	0.03	0.01	-0.02
Res-AV-CVAE-WithHM [8]	8.08	10.02	10.12	9.21	7.70	0.12	0.29	0.32	0.30	0.28	0.01	0.02	0.01	-0.01	-0.03
Res-AV-CVAE-DA-ST-GAN [8]	8.00	9.48	9.57	9.67	7.17	0.11	0.30	0.34	0.34	0.30	0	0.02	0.02	0	-0.02
Res-AV-CVAE-RFF [8]	9.62	10.72	10.68	9.70	8.00	0.22	0.45	0.46	0.43	0.35	0.03	0.05	0.05	0.01	-0.01
DualAVSE	16.06	15.21	14.09	12.98	11.27	0.35	0.54	0.74	0.92	1.01	0.10	0.10	0.08	0.05	0.03

Table 3: Results on MEAD datasets. Higher is better for all metrics.

References

- [1] Ahmed Hussen Abdelaziz et al. Ntcd-timit: A new database and baseline for noise-robust audio-visual speech recognition. In *Interspeech*, pages 3752–3756, 2017.
- [2] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third chime-

- speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511. IEEE, 2015.
- [3] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.
 - [4] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.
 - [5] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *Conference of the International Speech Communication Association*, Sep 2018.
 - [6] Ali Golmakani, Mostafa Sadeghi, and Romain Serizel. Audio-visual speech enhancement with a deep kalman filter generative model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
 - [7] Naomi Harte and Eoin Gillen. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Transactions on Multimedia*, 17(5):603–615, May 2015. ISSN 1520-9210, 1941-0077.
 - [8] Zhiqi Kang, Mostafa Sadeghi, Radu Horaud, Xavier Alameda-Pineda, Jacob Donley, and Anurag Kumar. The impact of removing head movements on audio-visual speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7302–7306. IEEE, 2022.
 - [9] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr-half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.
 - [10] Simon Leglaive, Laurent Girin, and Radu Horaud. A variance modeling framework based on variational autoencoders for speech enhancement. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.
 - [11] Karthik Ramesh, Chao Xing, Wupeng Wang, Dong Wang, and Xiao Chen. Vset: A multimodal transformer for visual speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6658–6662. IEEE, 2021.
 - [12] Mostafa Sadeghi and Xavier Alameda-Pineda. Mixture of inference networks for vae-based audio-visual speech enhancement. *IEEE Transactions on Signal Processing*, 69: 1899–1909, 2021.
 - [13] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Audio-visual speech enhancement using conditional variational autoencoders. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 28: 1788–1800, 2020.

- [14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [15] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments.
- [16] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020.
- [17] Wupeng Wang, Chao Xing, Dong Wang, Xiao Chen, and Fengyu Sun. A robust audio-visual speech enhancement model. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7529–7533. IEEE, 2020.
- [18] Haitao Xu, Liangfa Wei, Jie Zhang, Jianming Yang, Yannan Wang, Tian Gao, Xin Fang, and Lirong Dai. A multi-scale feature aggregation based lightweight network for audio-visual speech enhancement. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [19] Xinmeng Xu, Yang Wang, Jie Jia, Binbin Chen, and Dejun Li. Improving Visual Speech Enhancement Network by Learning Audio-visual Affinity with Multi-head Attention. In *Proc. Interspeech 2022*, pages 971–975, 2022.
- [20] Jing-Xuan Zhang, Genshun Wan, and Jia Pan. Is lip region-of-interest sufficient for lipreading? In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 368–372, 2022.
- [21] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 356–363. IEEE, 2020.