Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring

Supplementary Materials

1. Additional quantitative results

1.1. Previous methods on audio-visual corruption

We show additional results for the previous ASR, VSR, and AVSR models on audio-visual corrupted environments. Figure 2 shows the results on (a) visual occlusion corrupted environment, (b) visual noise corrupted environment, and (c) both visual occlusion and noise corrupted environment of each method. Similar to the results shown in Figure 3 in the manuscript, the previous AVSR models [1, 2] are not robust to the audio-visual corruption regardless of the visual corruption method. Note that Figure 2(c) is the same graph as Figure 3(a) in the original manuscript. In addition, we evaluate the robustness of the recent AVSR model, AV-HuBERT [3], trained using self-supervised learning. Table 1 shows the performance of AV-HuBERT on different audio and visual perturbations. When both streams are clean, the model achieves 1.48% WER which is a very high performance. However, when audio-visual corruption is introduced (*i.e.*, -5dB&Occ+Noise), the performance is degraded to 14.61% WER. The result indicates that even the recent state-of-the-art method also can be suffered performance degradation in audio-visual corruption situations. This is because the previous methods failed in considering modeling visual corruption during model training.

1.2. More results on audio-visual corruption

We compare the performances of the proposed method with baseline models using graphs. Figure 3 and Figure 4 show the comparison results on LRS2 and LRS3 datasets, respectively. We do not put the results of VSR model to focus on comparing with the previous AVSR and ASR models. In overall, the proposed AV-RelScore model outperforms the other baseline models and ASR model in different audio-visual corruption situations on both databases. The results confirm the effectiveness of the proposed method in improving the robustness to audio-visual input corruption.

1.3. More results on audio-only corruption

We additionally report the performance of the proposed AV-RelScore and the baseline models under an audio input-only corruption environment, the standard setting of previous methods [1, 2], on LRS3 dataset in Figure 1 (Figure 6 in the original manuscript reports the performance with the same setting using LRS2 dataset). By adopting the proposed audio-visual corruption modeling during training, all AVSR models achieve the better performances than the audio-only model, ASR. Moreover, by considering the reliability scores of each modal stream, the proposed AV-RelScore achieves the best performance on LRS3 as well as on LRS2. Through the results, we can confirm the importance of audio-visual corruption modeling in developing robust AVSR models and the effectiveness of the proposed AV-RelScore.

Audio Corruption & Video Corruption			
Clean&Clean	**-5dB&Clean**	**Clean&Occ+Noise**	**-5dB&Occ+Noise**
1.48%	11.18%	1.64%	14.61%

Table 1. WER of AV-HuBERT in different corruption situations.

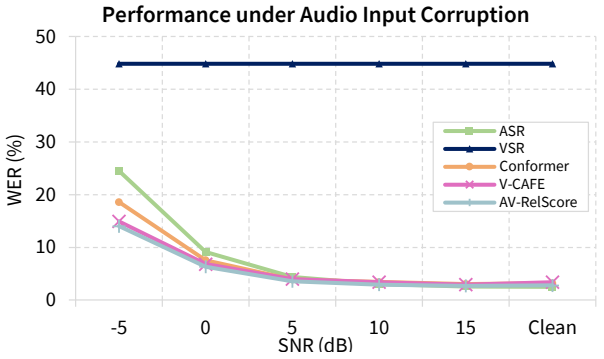

Figure 1. Speech recognition performances of ASR, VSR, and the proposed model under audio input-only corruption on LRS3. Note the models are trained with audio-visual corruption modeling.

2. Additional qualitative results

Figure 5 and Figure 6 show the additional visualization examples of the reliability scores of audio and visual modality from the LRS2 and LRS3 datasets, respectively. We exploit the same setting as in the original manuscript: the patch occlusion and noise for the visual corruption and noise with -5dB SNR for the audio corruption. From the vi-

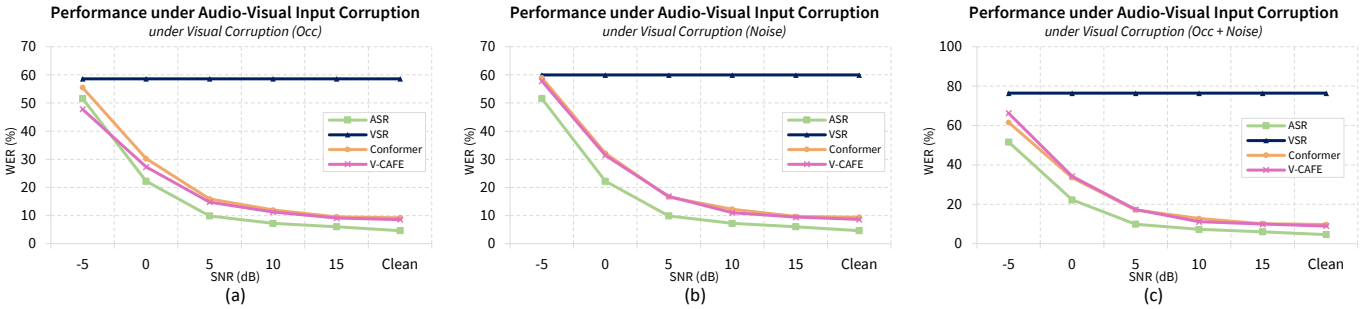

Figure 2. Speech recognition performances of ASR, VSR, and AVSR models under different audio-visual input corruption types on LRS2 dataset. (a) Under occlusion corruption. (b) Under visual noise corruption. (c) Under occlusion and noise corruption.

sual and audio reliability scores, we can clearly observe that the more highly corrupted visual inputs are, the less reliability scores are obtained, and vice versa. Also, each input modal well complements the other modal to recognize the speech properly. We additionally show the real-world occlusion case where the hands actually cover the lip of the subject on the bottom-right in Figure 6. In this case, we can clearly observe that the visual reliability scores are much less than those of other frames.

Furthermore, we provide a demo video including the reliability scores of the proposed AV-RelScore module and the actual prediction of our proposed model along with that of the recent state-of-the-art methodologies [1, 2] from both LRS2 and LRS3 datasets in ***demo_video.mp4***. We also include the case where the natural occlusion case appears in the dataset. Note that we will provide the code implementations for generating the corrupted audio-visual datasets for future release.

References

[1] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021. 1, 2

[2] Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro. Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition. *arXiv preprint arXiv:2207.06020*, 2022. 1, 2

[3] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 1

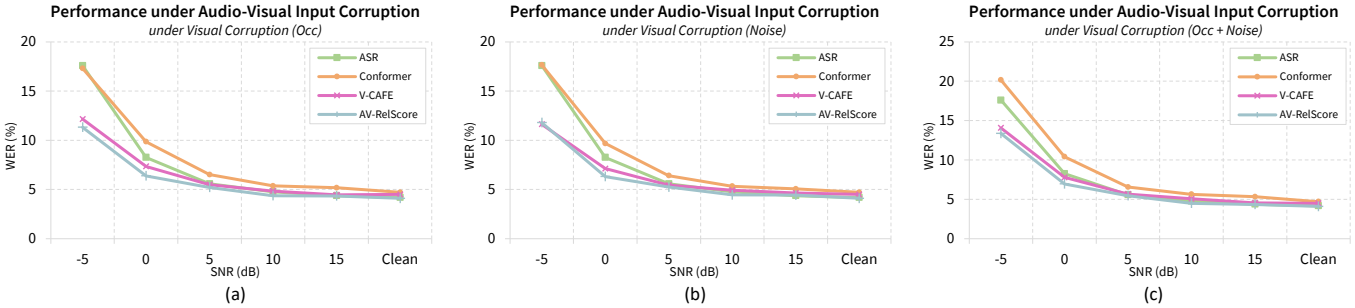

Figure 3. Speech recognition performances of ASR and AVSR models under different audio-visual input corruption types on LRS2 dataset. (a) Under occlusion corruption. (b) Under visual noise corruption. (c) Under occlusion and noise corruption.

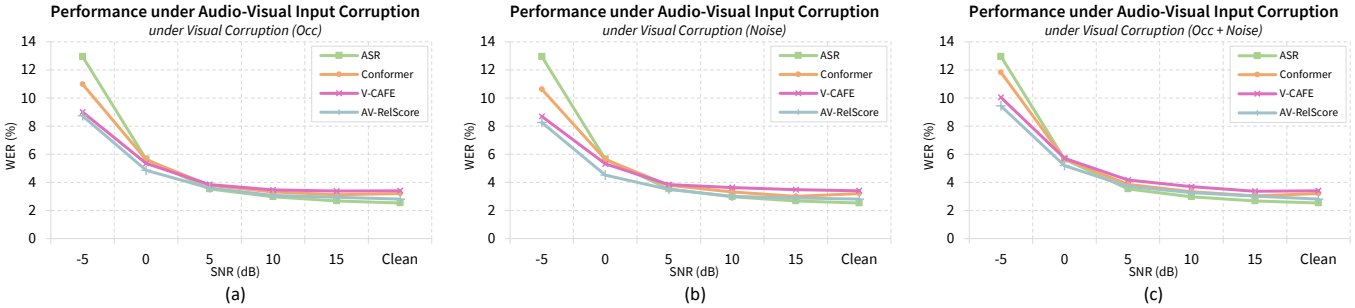

Figure 4. Speech recognition performances of ASR, and AVSR models under different audio-visual input corruption types on LRS3 dataset. (a) Under occlusion corruption. (b) Under visual noise corruption. (c) Under occlusion and noise corruption.

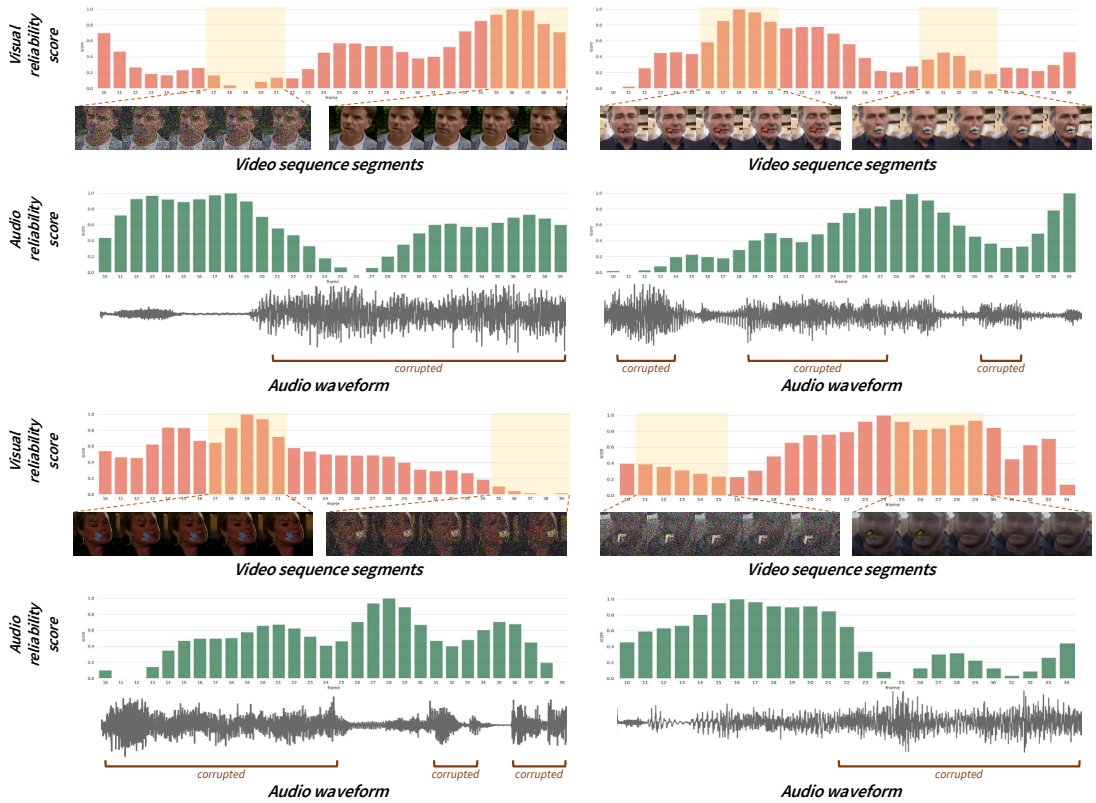

Figure 5. Visualization of visual reliability scores and audio reliability scores from AV-RelScore module of LRS2 dataset.

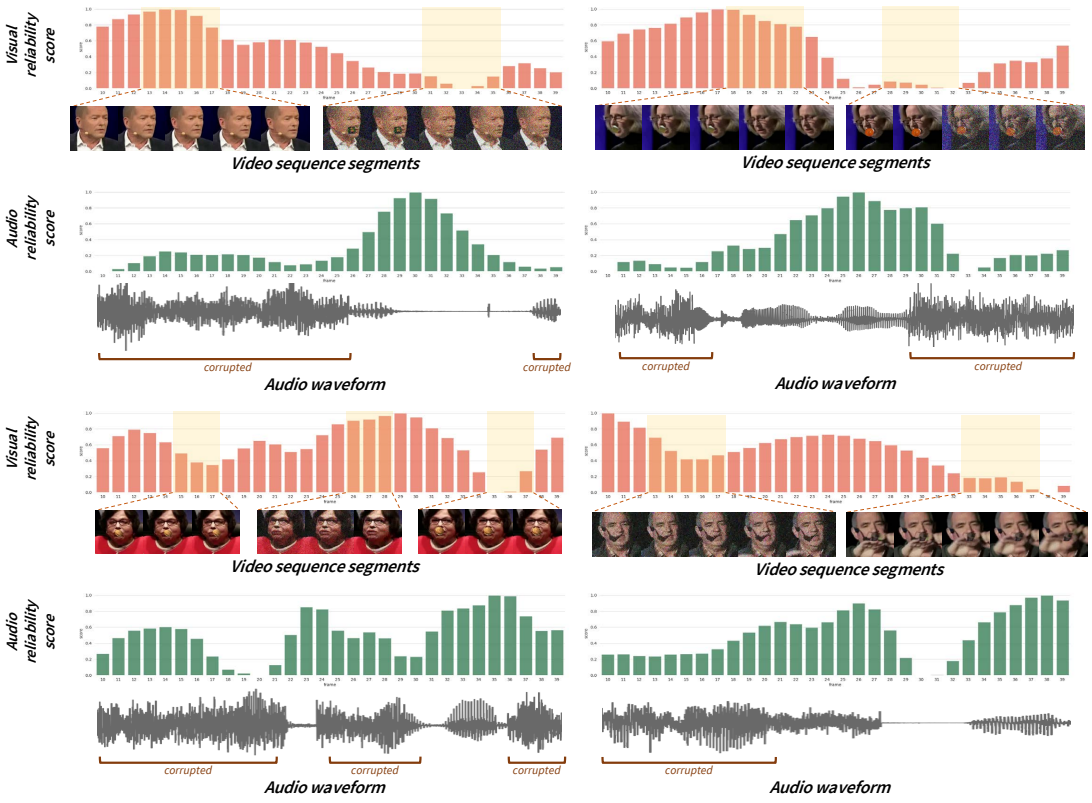

Figure 6. Visualization of visual reliability scores and audio reliability scores from AV-RelScore module of LRS3 dataset.