THU-AM-220

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

Joanna Hong[1*] Minsu Kim[1*] Jeongsoo Choi[1] Yong Man Ro[1†]

[1] Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

[*] equally contributed [†] Corresponding Author

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

KAIST

IMAGE & VIDEO SYSTEMS Lab. since 1998 Empowered by Deep Learning

Motivation

- Audio-visual speech recognition (AVSR) using audio and video data is nearly perfect and in the advanced stage.
- However, the previous studies have mostly considered the case where the audio inputs are corrupted and utilizing the additional clean visual inputs for complementing the corrupted audio information.
- Looking at the case, we come up with an important question, **what if both visual and audio information are corrupted, even simultaneously?**
- In real life, cases where both visual and audio inputs are corrupted alternatively or even simultaneously, are frequently happening.

Contribution

- We analyze the corruption of visual inputs using occlusion modeling in AVSR technique and design the robust training method in audio-visual speech recognition.
- We design a corrupted visual dataset inserting two types occlusions: occlusion patch and noise (i.e., blur and Gaussian).
- We propose Audio-Visual Reliability Scoring module (AV-RelScore) to figure out which modal is more reliable than other to recognize the input speech when either of one is corrupted, or even both.
- We conduct comprehensive experiments with all ASR, VSR, and AVSR task to validate the effectiveness of the proposed task modeling and network architecture with LRS2 and LRS3 the largest audio-visual datasets obtained in the wild.

Analysis

- **Robustness of AVSR to acoustic and visual noise**

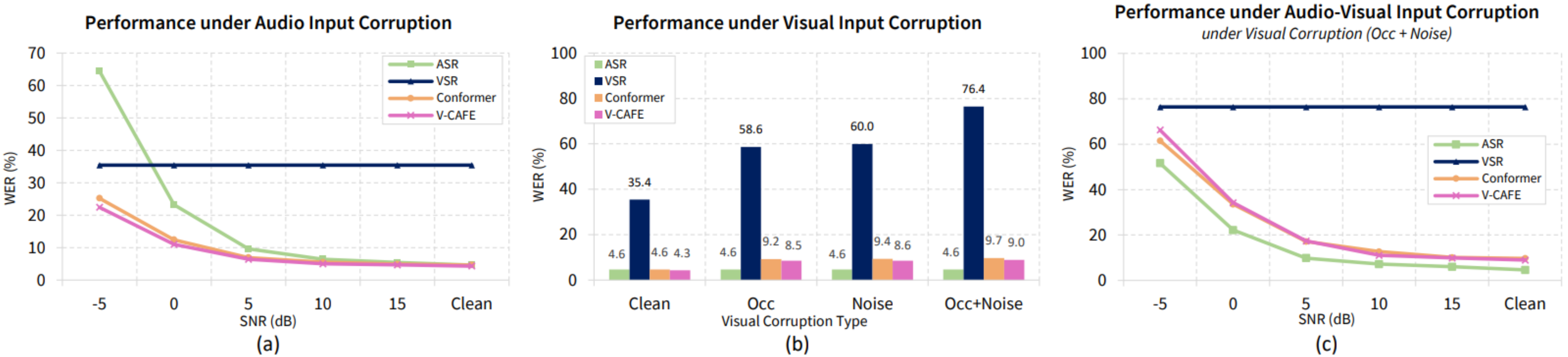


Figure 3. Speech recognition performances of ASR, VSR, and AVSR models on LRS2 dataset under different input corruption types: (a) Audio input corruption with babble noise. (b) Visual input corruption with occlusion and noise. (c) Audio-visual input corruption.

Data construction

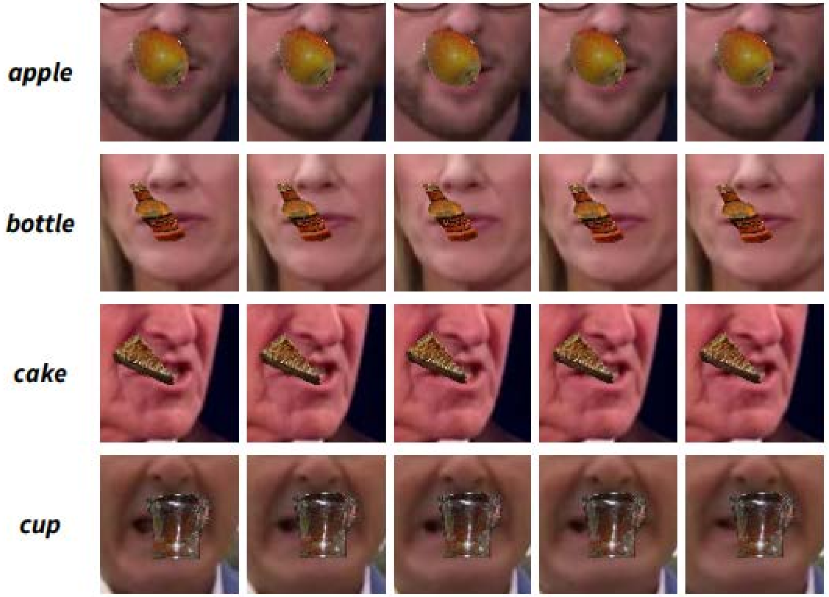


Figure 1. Examples of visual occlusion with NatOcc patches.

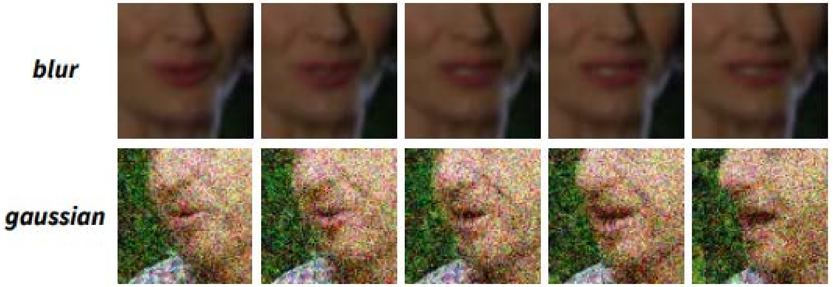


Figure 2. Examples of visual corruption with noises.

- **Visual occlusion with patches**
  - Given the input lip-centered talking face video, we randomly choose how many times the occlusion occurs in whole sequences.
  - Then, we randomly select the video frames that we are going to attach the patch and put the random patch along with the random position of lip landmarks.
- **Visual corruption with noises**
  - We randomly insert blur or gaussian noise to the entire input face video, respectively. Otherwise, we utilize the clean sequence.
  - For visual corruption with noise, we also follow the same scheme of selecting the random ratio t applied in the visual corruption with occlusion patches.

Proposed Architecture

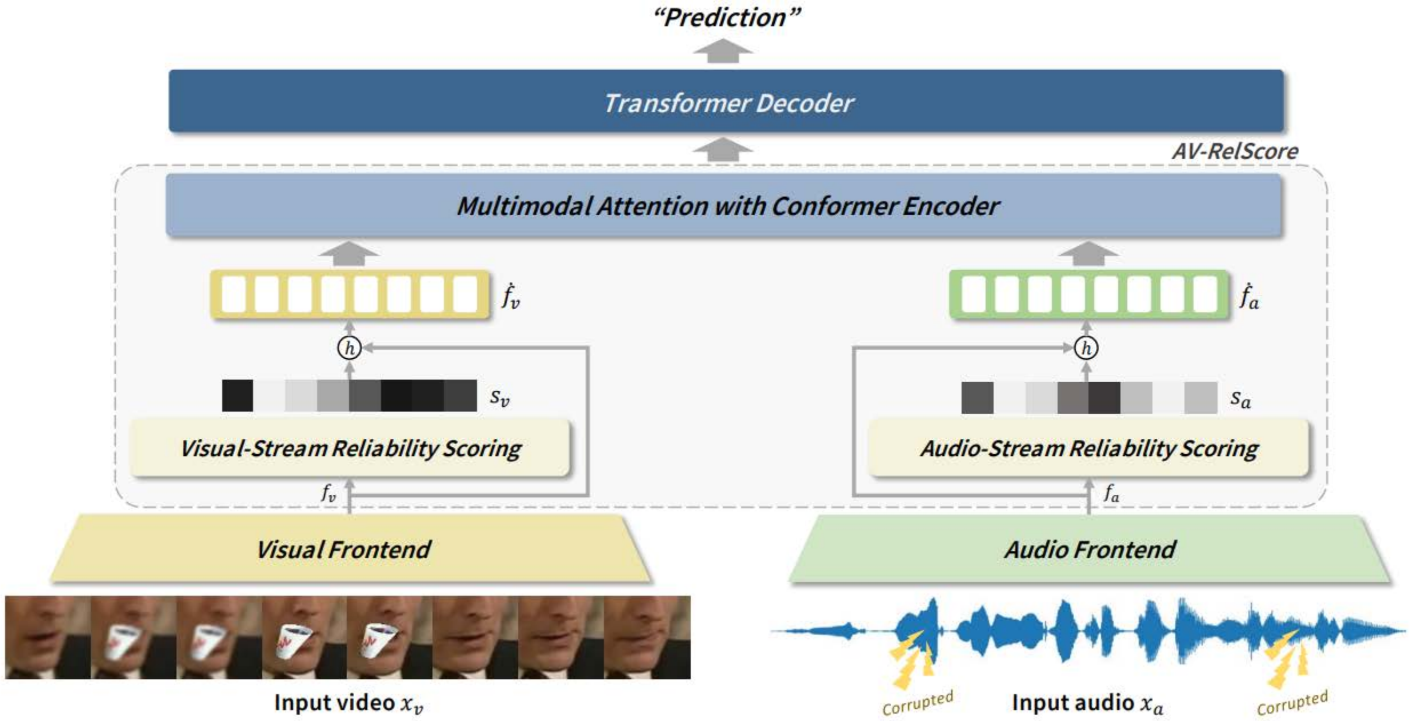


Figure 4. Overall architecture of the proposed AVSR framework.

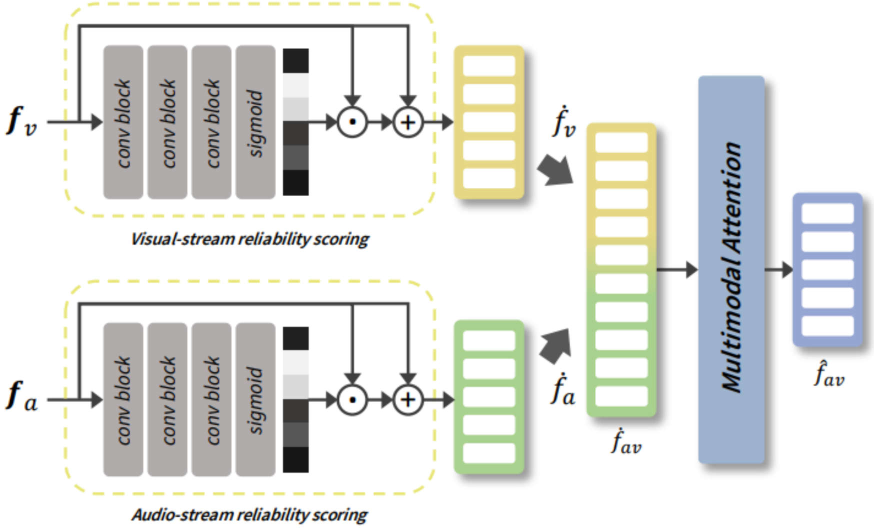
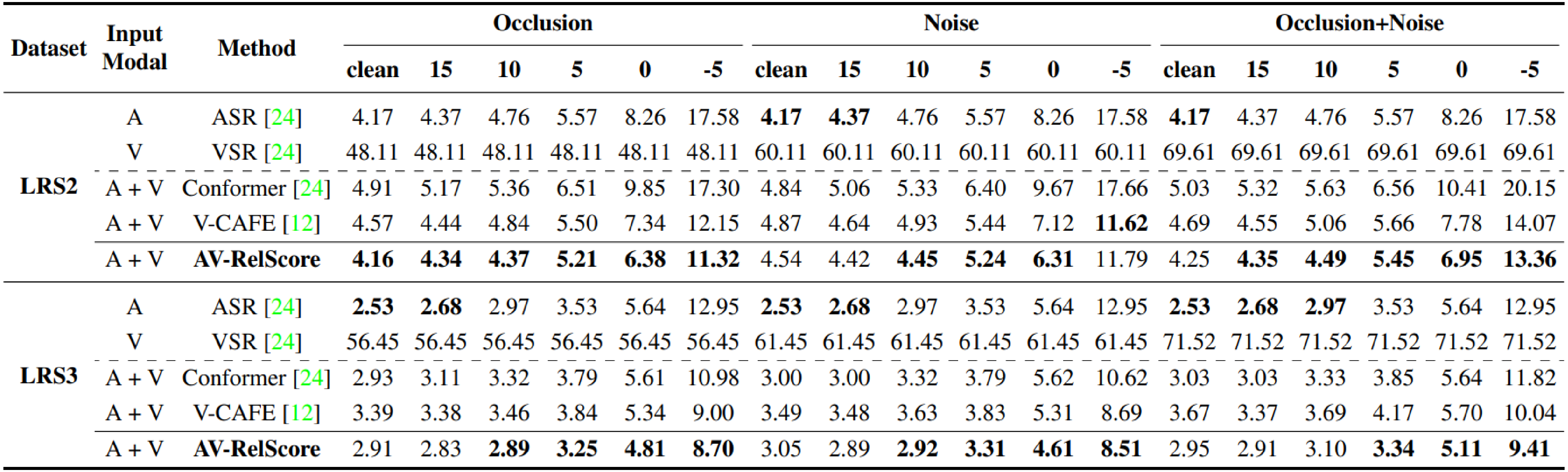


Figure 5. Detailed architecture of AV-RelScore.

Experiment

- **Dataset**
  - **LRS2**
    - An English sentence-level audio-visual dataset that collected from BBC television shows. It has about 142,000 utterances including pre-train and train sets, about 1,000 utterances for validation set, and about 1,200 utterances for test set. We utilize both sets for training, and test the model on a test set containing 1,243 utterances.
  - **LRS3**
    - A large-scale English sentence-level audio-visual dataset. It consists of about 150,000 videos which are total about 439 hours long and collected from TED. About 131,000 utterances are utilized for training, and about 1,300 utterances are used for testing.

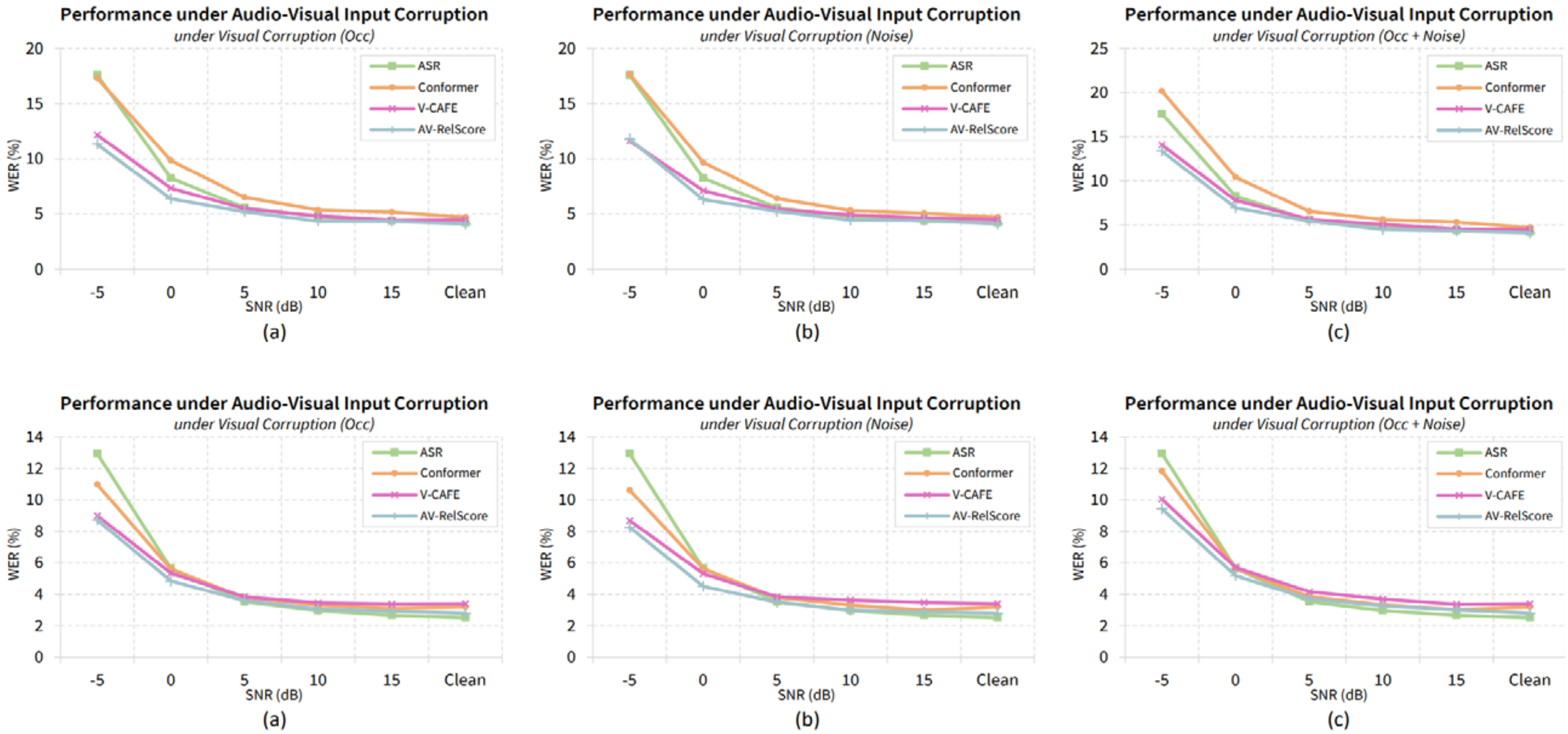
Experimental Result



Dataset	Input Modal	Method	Occlusion						Noise						Occlusion+Noise					
			clean	15	10	5	0	-5	clean	15	10	5	0	-5	clean	15	10	5	0	-5
LRS2	A	ASR [24]	4.17	4.37	4.76	5.57	8.26	17.58	**4.17**	**4.37**	4.76	5.57	8.26	17.58	**4.17**	4.37	4.76	5.57	8.26	17.58
	V	VSR [24]	48.11	48.11	48.11	48.11	48.11	48.11	60.11	60.11	60.11	60.11	60.11	60.11	69.61	69.61	69.61	69.61	69.61	69.61
	A + V	Conformer [24]	4.91	5.17	5.36	6.51	9.85	17.30	4.84	5.06	5.33	6.40	9.67	17.66	5.03	5.32	5.63	6.56	10.41	20.15
	A + V	V-CAFE [12]	4.57	4.44	4.84	5.50	7.34	12.15	4.87	4.64	4.93	5.44	7.12	**11.62**	4.69	4.55	5.06	5.66	7.78	14.07
	A + V	**AV-RelScore**	**4.16**	**4.34**	**4.37**	**5.21**	**6.38**	**11.32**	4.54	4.42	**4.45**	**5.24**	**6.31**	11.79	4.25	**4.35**	**4.49**	**5.45**	**6.95**	**13.36**
LRS3	A	ASR [24]	**2.53**	**2.68**	2.97	3.53	5.64	12.95	**2.53**	**2.68**	2.97	3.53	5.64	12.95	**2.53**	**2.68**	**2.97**	3.53	5.64	12.95
	V	VSR [24]	56.45	56.45	56.45	56.45	56.45	56.45	61.45	61.45	61.45	61.45	61.45	61.45	71.52	71.52	71.52	71.52	71.52	71.52
	A + V	Conformer [24]	2.93	3.11	3.32	3.79	5.61	10.98	3.00	3.00	3.32	3.79	5.62	10.62	3.03	3.03	3.33	3.85	5.64	11.82
	A + V	V-CAFE [12]	3.39	3.38	3.46	3.84	5.34	9.00	3.49	3.48	3.63	3.83	5.31	8.69	3.67	3.37	3.69	4.17	5.70	10.04
	A + V	**AV-RelScore**	2.91	2.83	**2.89**	**3.25**	**4.81**	**8.70**	3.05	2.89	**2.92**	**3.31**	**4.61**	**8.51**	2.95	2.91	3.10	**3.34**	**5.11**	**9.41**

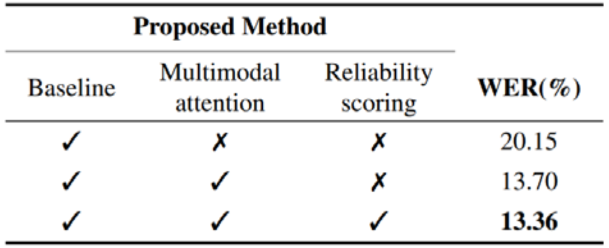
Table 1. WER (%) comparisons with the state-of-the-art methods on audio-visual corrupted environment. The first row represents the types of visual corruption: patch occlusion, noise, and both, and the second row indicates audio noise with different levels, SNR(dB).

Experimental Result



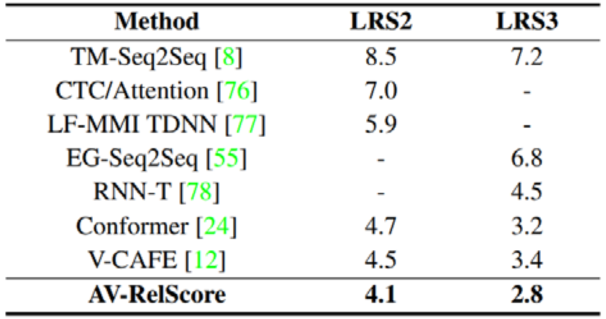
(a) (b) (c)

(a) (b) (c)



Proposed Method			WER(%)
Baseline	Multimodal attention	Reliability scoring	
✓	✗	✗	20.15
✓	✓	✗	13.70
✓	✓	✓	**13.36**

Table 2. Ablation study on LRS2 dataset.



Method	LRS2	LRS3
TM-Seq2Seq [8]	8.5	7.2
CTC/Attention [76]	7.0	-
LF-MMI TDNN [77]	5.9	-
EG-Seq2Seq [55]	-	6.8
RNN-T [78]	-	4.5
Conformer [24]	4.7	3.2
V-CAFE [12]	4.5	3.4
AV-RelScore	**4.1**	**2.8**

Table 3. WER (%) comparisons with state-of-the-art methods.

Experimental Result

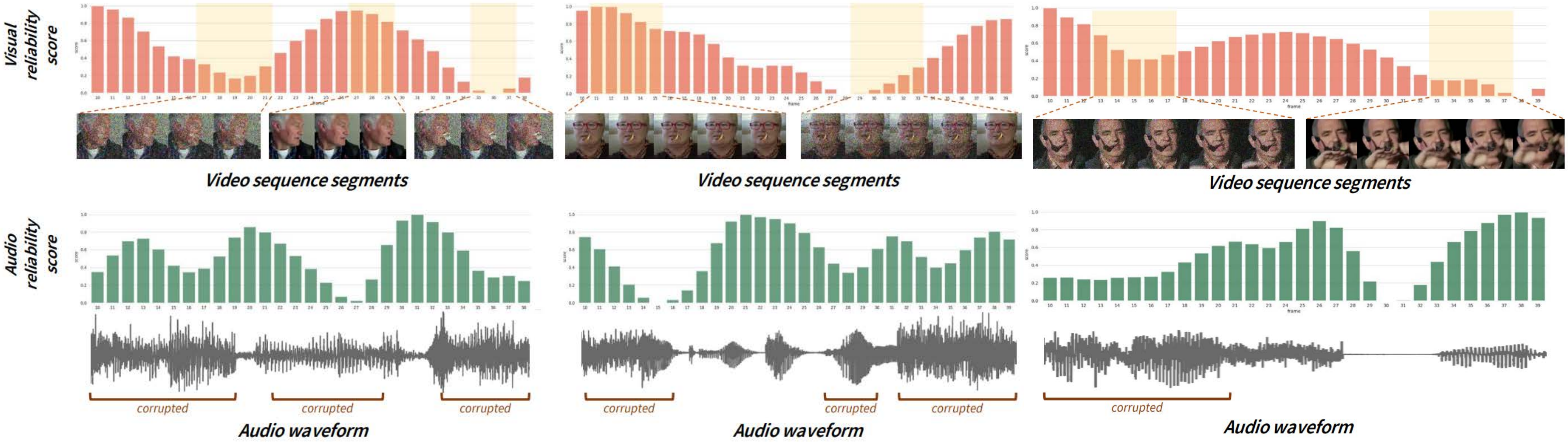


Figure 7. Visualization of visual reliability scores and audio reliability scores from AV-RelScore module

Thank you!