SUPPLEMENT FOR PAPER ID 3956
UNIFIED STREAMING AND NON-STREAMING MODEL FOR AUDIO-VISUAL SPEECH RECOGNITION

Table 1. WER (%) of the proposed streaming and non-streaming models in babble noise environments. The chunk size is set to 16 for streaming models which means that the latency of streaming model is 400ms.

Modality	Streaming	SNR					Clean
		-5dB	0dB	5dB	10dB	15dB	
Audio-only	N	65.2	21.8	7.7	4.1	3.1	2.4
	Y	65.9	25.1	10.1	6.0	4.4	3.4
Audio-visual	N	25.9	10.6	4.9	3.3	2.7	2.3
	Y	33.1	14.5	7.4	5.1	4.3	3.6

As shown in Table 1, the proposed streaming audio-visual model can get better results than the streaming audio-only model. It shows the obvious advantage of our unified audio-visual model in the streaming mode. The audio-only model also achieves the SOTA result in the ASR benchmark.

Table 2. Ablation study of the Ro-conformer encoder. ✓ denotes that the Ro-conformer is used in the corresponding module, while × denotes that the Ro-conformer is not used. The evaluation metric is WER(%).

Audio back-end	Visual back-end	Fusion encoder	Test set
✓	×	✓	2.3
✓	✓	✓	2.5
✓	✓	×	2.6

As shown in the Table 2, the audio-visual model using Ro-Conformer in the visual back-end got a WER of 2.5%, while the audio-visual model that does not use Ro-Conformer in the visual back-end got a WER of 2.3%. On other hand, the audio-visual model that does not use Ro-Conformer in the fusion encoder got a WER of 2.6%. The experimental results show that Ro-Conformer is not suitable for modeling video features.