Watch to Listen Clearly: Visual Speech Enhancement Driven Multi-modality Speech Recognition
Supplementary Material

Bo Xu, Jacob Wang, Cheng Lu and Yandong Guo
Xpeng motors
xiaoboboer@gmail.com

1. Architecture details of the VE sub-network

Table 1 and Table 2 show the architecture details of the visual speech enhancement (VE) sub-network.

Layer	# filters	K	S	P	Out
fc0	1536	1	1	1	T
conv1	1536	5	1	2	T
conv2	1536	5	1	2	T
conv3	1536	5	$\frac{1}{2}$	2	$2T$
conv4	1536	5	1	2	$2T$
conv5	1536	5	1	2	$2T$
conv6	1536	5	1	2	$2T$
conv7	1536	5	$\frac{1}{2}$	2	$4T$
conv8	1536	5	1	2	$4T$
conv9	1536	5	1	2	$4T$
fc10	256	1	1	1	$4T$

(a) Video Stream.

Layer	# filters	K	S	P	Out
fc0	1536	1	1	1	$4T$
conv1	1536	5	1	2	$4T$
conv2	1536	5	1	2	$4T$
conv3	1536	5	1	2	$4T$
conv4	1536	5	1	2	$4T$
conv5	1536	5	1	2	$4T$
fc6	256	1	1	1	$4T$

(b) Noisy audio Stream.

Table 1: Architecture details of the visual speech enhancement (VE) sub-network (*Part I*). **a)** The 1D ResNet module of video stream that extracts the video features. **b)** The 1D ResNet module of audio stream that extracts the noisy audio features. **K:** Kernel width; **S:** Stride – fractional strides denote transposed convolutions; **P:** Padding; **Out:** Temporal dimension of the layers output.

Layer	# filters	Out
EleAtt-GRU	512	$4T$
fc1	600	$4T$
fc2	600	$4T$
fc_mask	F	$4T$

(a) AV Fusion.

Layer	# filters	K	S	P	Out
fc0	1536	1	1	1	$4T$
conv1	1536	5	2	2	$2T$
EleAtt-GRU	128	-	-	-	$2T$
conv2	1536	5	2	2	T
fc6	512	1	1	1	T

(b) Enhanced audio stream.

Table 2: Architecture details of the AE sub-network (*Part II*). **a)** The EleAtt-GRU and FC layers that process multimodality fusion and enhancing encoding. **b)** The EleAtt-GRU and 1D ResNet layers that extracts the enhanced audio features.