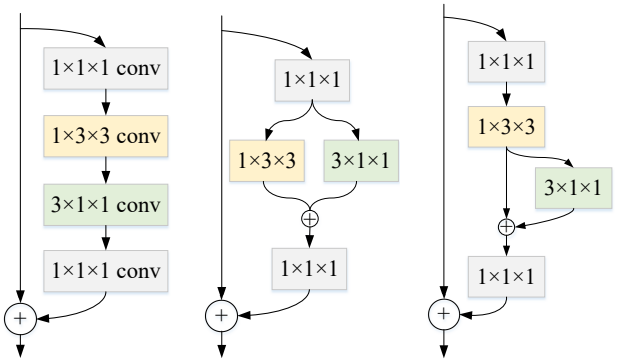
Discriminative Multi-modality Speech Recognition
Supplementary Material

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

The following supplementary material includes the details: **1)** P3D network (Sec. 1); **2)** EleAtt-GRU block (Sec. 2); **3)** examples of AE and AE-MSR speech recognition results (Sec. 3); **4)** enhancement examples of the AE networks (Sec. 4); **5)** examples of mouth crop (Sec. 5); **6)** architecture of the AE networks (Sec. 6).

1. Blocks of the Pseudo-3D (P3D) network

P3D ResNet is implemented by separating $N \times N \times N$ convolutions into $1 \times 3 \times 3$ convolution filters on spatial domain and $3 \times 1 \times 1$ convolution filters on temporal domain to extract spatial-temporal features [2]. The three versions of P3D blocks are shown in Figure 1.



(a) P3D-A (b) P3D-B (c) P3D-C

Figure 1: Bottleneck building blocks of the Pseudo-3D (P3D) ResNet network. P3D ResNet is produced by interleaving P3D-A, P3D-B and P3D-C in turn.

2. Details of an EleAtt-GRU block

The details of the EleAtt-GRU [3] building block used by our models are outlined in Figure 2. Each GRU block has (*e.g.*, N) GRU neurons. **Yellow** boxes – the units of the original GRU with the output dimension of N. **Blue circle** – element-wise operation and the brown circle denotes vector addition operation. **Red box** – EleAttG with an output dimension of D, which is the same as the dimension of the input x_t.

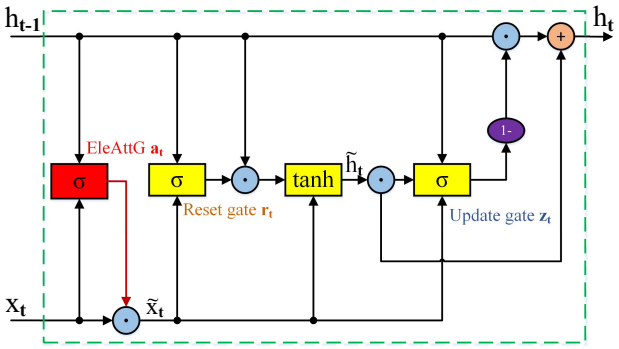


Figure 2: An Element-wise-Attention Gate (EleAttG) of GRU block.

Corresponding computations of an EleAtt-GRU are as follows:

$$\begin{aligned}
\tilde{x_t} &= a_t \odot x_t \\
r_t &= \sigma\left(W_{xr}\tilde{x_t} + W_{hr}h_{t-1} + b_r\right) \\
z_t &= \sigma\left(W_{xz}\tilde{x_t} + W_{hz}h_{t-1} + b_z\right) \\
h_t{}' &= \tanh\left(W_{xh}\tilde{x_t} + W_{hh}\left(r_t \odot h_{t-1}\right) + b_h\right) \\
h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot h_t{}'
\end{aligned}$$

where σ denotes the activation function of Sigmoid. The attention response of an EleAttG is the vector a_t with the same dimension as the input x_t of GRU. a_t modulates x_t to generate $\tilde{x_t}$. r_t and z_t denote the reset gate and update gate of GRU. h_t and h_{t-1} are the output vectors of the current hidden state and the previous hidden state. $W_{\alpha\beta}$ denotes the weight matrix related with α and β, where $\alpha \in \{x, h\}$ and $\beta \in \{r, z, h\}$. b_γ is the bias vector, where $\gamma \in \{r, z, h\}$ [3].

3. Examples of AE and AE-MSR speech recognition results.

Examples of AE and AE-MSR speech recognition results are illustrated in Table 1.

	Transcription	ΔM%	WER %
GT	We can prevent the worst case scenario	-	
V	We can put and worst case scenario	-	34
A	We can prevent the worst case tcario	-	8
AV	We can prevent the worst case scenario	-	0
	Noisy (5 dB)		
GT	what would that change about how we live	-	
V	wouldn't at chance whole a life	-	53
A	that would I try all we live	36	50
AV	that would I chance all how we live	24	25
VA (1DRN)	what would that change about how we live	11	0
	Noisy (0 dB)		
GT	human relationships are not efficient	-	
V	you man relation share are now efficient	-	38
A	man went left now fit	80	73
AV	you man today are now efficient	89	43
VA (1DRN)	human relations are now efficient	31	14
VAV (1DRN)	human relationships are not efficient	21	0
	Noisy (0 dB)		
GT	we really don't walk anymore	-	
V	we aren't working	-	61
A	wh ae lly son't tank	63	50
AV	we alley won't work more	63	32
VA (1DRN)	we really won't work anymore	39	11
VA (TCN)	we really don't walk anymore	22	0
	Noisy (-5 dB)		
GT	at some point I'm going to get out	-	
V	I soon planning to get it	-	47
A	it so etolunt	96	76
AV	at soon pant talking to get it	96	35
VAV (1DRN)	at some point I'm taking to get out	33	9
VAV (TCN)	at some point I'm going to get out	20	0

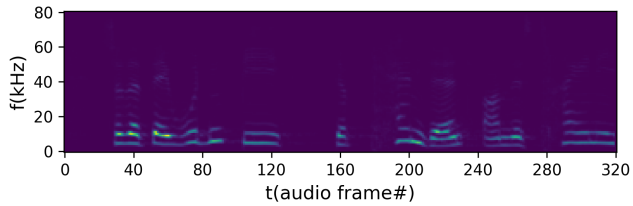
Table 1: Examples of recognition results by our models. **GT:** Ground truth; **V:** visual modality only; **A:** audio modality only; **AV:** multi-modality with single visual modality awareness; **VA:** enhanced audio modality by single visual awareness for ASR; **VAV:** multi-modality by double visual awareness for multi-modality speech recognition (MSR); **1DRN, TCN:** the temporal convolutional unit is 1D ResNet or TCN.

4. Enhancement examples of the 1DRN-AE and the TCN-AE models

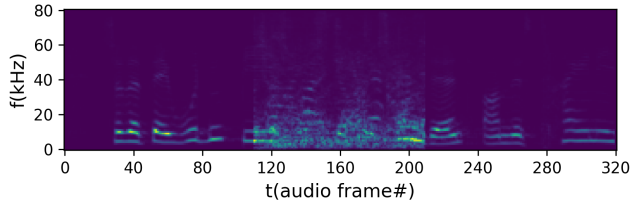
Enhancement examples of our audio enhancement sub-networks are illustrated in Figure 3.

5. Examples of mouth crop

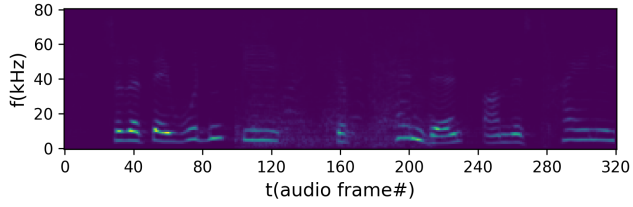
We produce image frames by cropping original video frames to 112×112 pixel patches and choose mouth patch as region of interest (ROI). As shown in Figure 4, facial landmarks are extracted by the *Dlib* [1] toolkit and the mouth ROI inside the red squares are achieved by 4 (red points) specified out of 68 facial landmarks (green points).



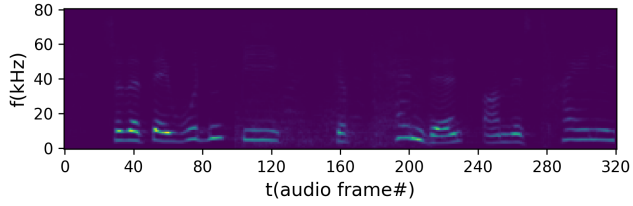
(a) Clean audio magnitude



(b) Noisy audio magnitude



(c) Enhanced audio magnitude by 1DRN-AE



(d) Enhanced audio magnitude by TCN-AE

Figure 3: Enhancement effects of the 1D-ResNet-based audio enhancement (1DRN-AE) model and the TCN-based audio enhancement (TCN-AE) model: **a)** clean audio utterance; **b)** we obtain this noisy utterance by adding babble noise to the 100 central audio frames; **c)** the enhanced audio utterance by 1DRN-AE; **d)** the enhanced audio utterance by TCN-AE; **c)** and **d)** show the effect of audio enhancement when compared to **b)**.

6. Architecture details of the audio enhancement networks

Architecture details of the audio enhancement sub-network are given in Table 2.

Figure 4: Examples of mouth crop.

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 of 1D ResNet.

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) Audio Stream of 1D ResNet.

Layer	Hidden	K	N	S	Out
fc0	520	1	1	1	$4T$
TCN1	520	3	3	1	$4T$
fc2	256	1	1	1	$4T$

(c) Audio stream of TCN.

Layer	Hidden	K	N	S	Out
fc0	520	1	1	1	T
TCN1	520	3	3	1	T
conv2	520	3	1	$\frac{1}{2}$	$2T$
TCN3	520	3	3	1	T
conv4	520	3	1	$\frac{1}{2}$	$4T$
fc5	256	1	1	1	$4T$

(d) Video Stream of TCN.

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

(e) 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

(f) Enhanced audio stream.

Table 2: Architecture details. **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. **c)** The TCN module of video stream that extracts the video features. **d)** The TCN module of audio stream that extracts the noisy audio features. **e)** The EleAtt-GRU and FC layers that process multi-modality fusion and enhancing encoding. **f)** The EleAtt-GRU and 1D ResNet layers that extracts the enhanced audio features. **K:** Kernel width; **S:** Stride – fractional strides denote transposed convolutions; **P:** Padding; **Out:** Temporal dimension of the layer's output. **Hidden:** the number of hidden units; **N:** the number of TCN blocks.

References

[1] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[2] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. *IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[3] Pengfei Zhang, Jianru Xue, Cuiling Lan, Wenjun Zeng, Zhanning Gao, and Nanning Zheng. Eleatt-rnn: Adding attentiveness to neurons in recurrent neural networks. *arXiv preprint arXiv:1909.01939*, 2019.