Multi-grained Spatio-Temporal Features Perceived Network for Event-based Lip-Reading

Supplementary Material

In this part, we first describe more details about the vocabulary selection process of the DVS-Lip dataset. Then, we present more qualitative visualization results on the test set of the DVS-Lip dataset.

A. Vocabulary Selection

To explore the advantages of event cameras in capturing fine-grained movement evolution information, we divide the vocabulary of the DVS-Lip dataset into two parts, where the first part is composed of visually similar word pairs and the second part is composed of common words. The first part of the vocabulary consists of the 25 most frequently confused word pairs that are selected from the vocabulary (500 words in total) of the LRW dataset [1]. To evaluate the confusions between words, we first run the model in [2] on the test set of the LRW dataset. Then, the confusion of each word is set to the proportion of the most common incorrect prediction results corresponding to the word. This is consistent with the word pair confusion evaluation in [1]. The most frequently confused word pairs are shown in Table 1. For the second part of our vocabulary, we randomly select another 50 words from the vocabulary of the LRW dataset. Combining the two parts, the vocabulary of the DVS-Lip dataset contains a total of 100 words. The full list of the vocabulary is shown in Table 2.

Label	Prediction	Proportion	Label	Prediction	Proportion
price	press	0.20	happened	happen	0.12
difference	different	0.18	Syrian	Syria	0.12
benefits	benefit	0.16	taking	taken	0.12
little	legal	0.16	challenge	change	0.12
million	billion	0.16	terms	times	0.10
worst	words	0.16	around	ground	0.10
spend	spent	0.16	missing	meeting	0.10
think	thing	0.16	called	court	0.10
number	numbers	0.14	election	action	0.10
allow	allowed	0.14	giving	evening	0.10
American	America	0.14	paying	being	0.10
heavy	having	0.14	these	needs	0.10
Russian	Russia	0.14			

Table 1. Labels and their corresponding most frequently mispredictions, results come from the model in [2] on LRW dataset [1].

B. Qualitative Results

In this section, we present more qualitative results by applying the Grad-CAM [3] to our MSTP using the samples from the DVS-Lip test set. Figure 1 and Figure 2 show the examples from the first part of the test set, and Figure 3 shows the examples from the second part of the test set. The first row of each example shows the saliency maps for the low-rate branch's input event frames ($T^{low} = 30$), and the second row shows the saliency maps for the input event frames ($T^{high} = 210$) of the high-rate branch. The saliency maps for the high-rate branch are downsampled by a factor of 7 so that the saliency maps from the two branches are aligned in time. We only display the event frames that contain the corresponding word. These results demonstrate that our MSTP can automatically select important spatio-temporal regions for word recognition from event frame inputs of different granularities. Accordingly, the model can learn both complete spatial features and fine temporal features.

Part1	allow	allowed	America	American	benefit
	benefits	challenge	change	court	called
	different	difference	happen	happened	heavy
	having	little	legal	million	billion
	number	numbers	price	press	Syria
	Syrian	taking	taken	think	thing
	worst	words	around	ground	terms
	times	paying	being	missing	meeting
	election	action	giving	evening	Russia
	Russian	spend	spent	these	needs
Part2	tomorrow	right	still	years	significant
	become	house	everything	should	warning
	economic	several	young	majority	attacks
	exactly	accused	death	hundreds	support
	described	labour	chief	welcome	leaders
	water	during	under	England	judge
	general	saying	between	capital	started
	security	perhaps	minutes	potential	another
	couple	banks	Germany	point	London
	immigration	question	really	military	education

Table 2. Vocabulary of the DVS-Lip dataset.

References

[1] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016. 1

[2] Dalu Feng, Shuang Yang, Shiguang Shan, and Xilin Chen. Learn an effective lip reading model without pains. *arXiv preprint arXiv:2011.07557*, 2020. 1

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

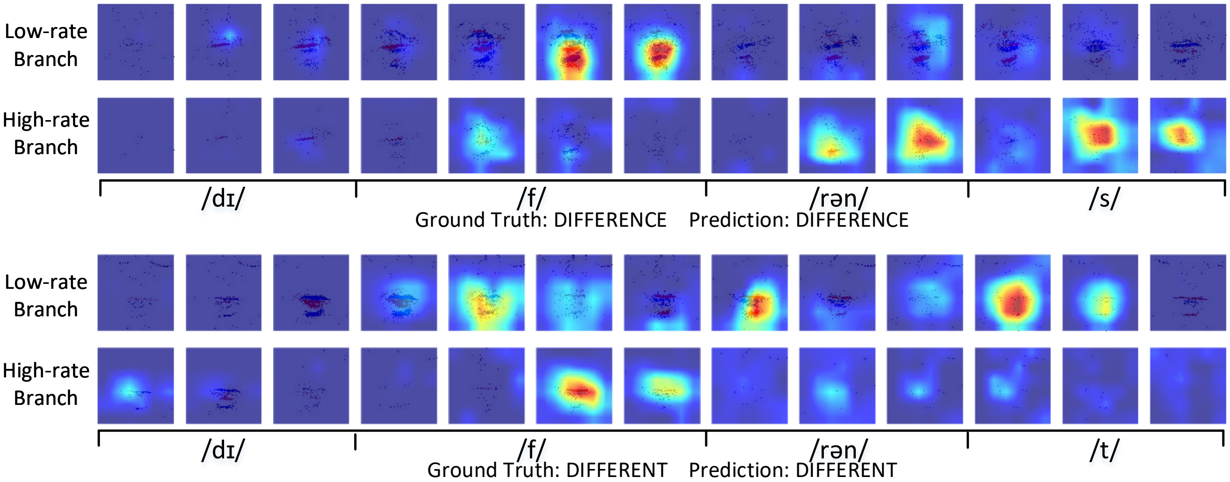

Figure 1. Visualization of the saliency maps for words “difference” and “different”.

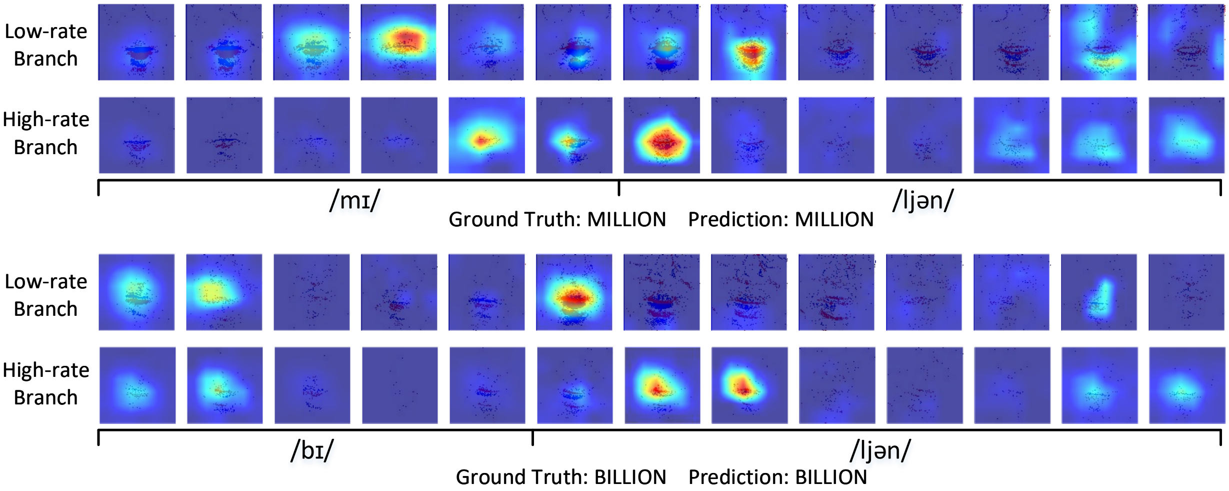

Figure 2. Visualization of the saliency maps for words “million” and “billion”.

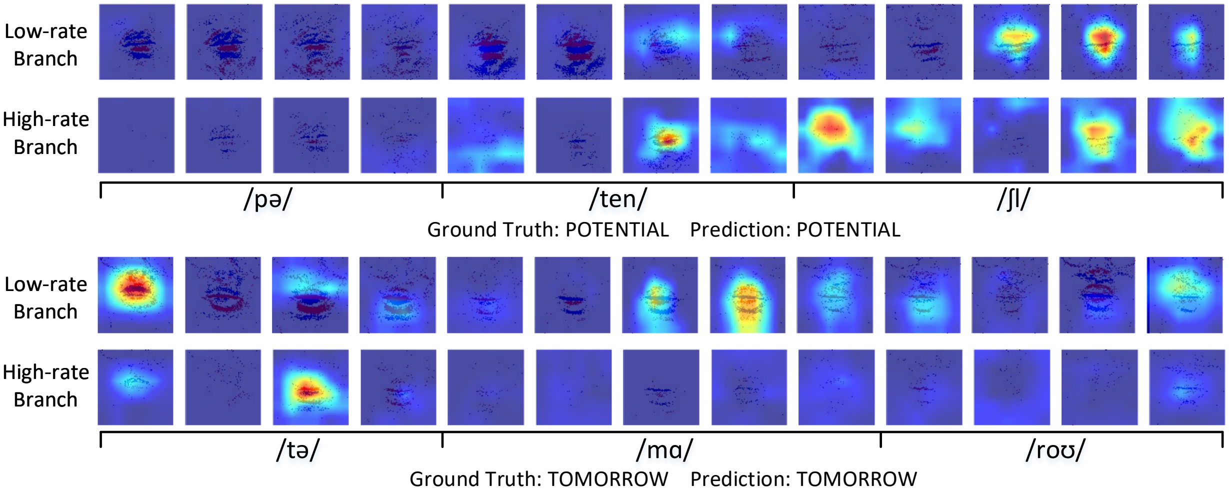

Figure 3. Visualization of the saliency maps for words “potential” and “tomorrow”.