Supplementary Material

Kyle Min[1]*, Sourya Roy[2]*†, Subarna Tripathi[1], Tanaya Guha[3], and Somdeb Majumdar[1]

[1]Intel Labs, [2]UC Riverside, [3]University of Glasgow

A Network architecture details

For clarification, network architecture of SPELL is shown in Table 1.

Table 1. Detailed architecture of SPELL. Batch normalization [2] and ReLU [3] follow after each of (6), (7), and (9-14).

Index	Inputs	Description	Dimension
(1)	-	4-D spatial feature	4
(2)	-	Visual feature v_{visual}	512
(3)	-	Audio feature v_{audio}	512
(4)	(1)	Linear ($4 \rightarrow 64$)	64
(5)	(2), (4)	Concatenation	576
(6)	(5)	Linear ($576 \rightarrow 64$)	64
(7)	(3)	Linear ($512 \rightarrow 64$)	64
(8)	(6), (7)	Addition	64
(9)	(8)	EDGE-CONV (Forward)	64
(10)	(8)	EDGE-CONV (Undirected)	64
(11)	(8)	EDGE-CONV (Backward)	64
(12)	(9)	SAGE-CONV (Forward, Shared)	64
(13)	(10)	SAGE-CONV (Undirected, Shared)	64
(14)	(11)	SAGE-CONV (Backward, Shared)	64
(15)	(12)	SAGE-CONV (Forward)	1
(16)	(13)	SAGE-CONV (Undirected)	1
(17)	(14)	SAGE-CONV (Backward)	1
(18)	(15), (16), (17)	Addition	1
(19)	(18)	Sigmoid	1

B Qualitative results

In Figure 1, we provide qualitative results of SPELL. For additional comparison, we also show the results of ASC [1], which performs the best among the approaches that have released their model weights as well as the source code. In the first example, ASC has both false positive and false negative results while SPELL is able to correctly classify all the speakers. In the second and the third examples, SPELL finds every active speaker when ASC has many false negatives. The results show that SPELL is effective and is good at modeling spatial-temporal long-term information. For more examples, please refer to the videos that are included in the subfolder.

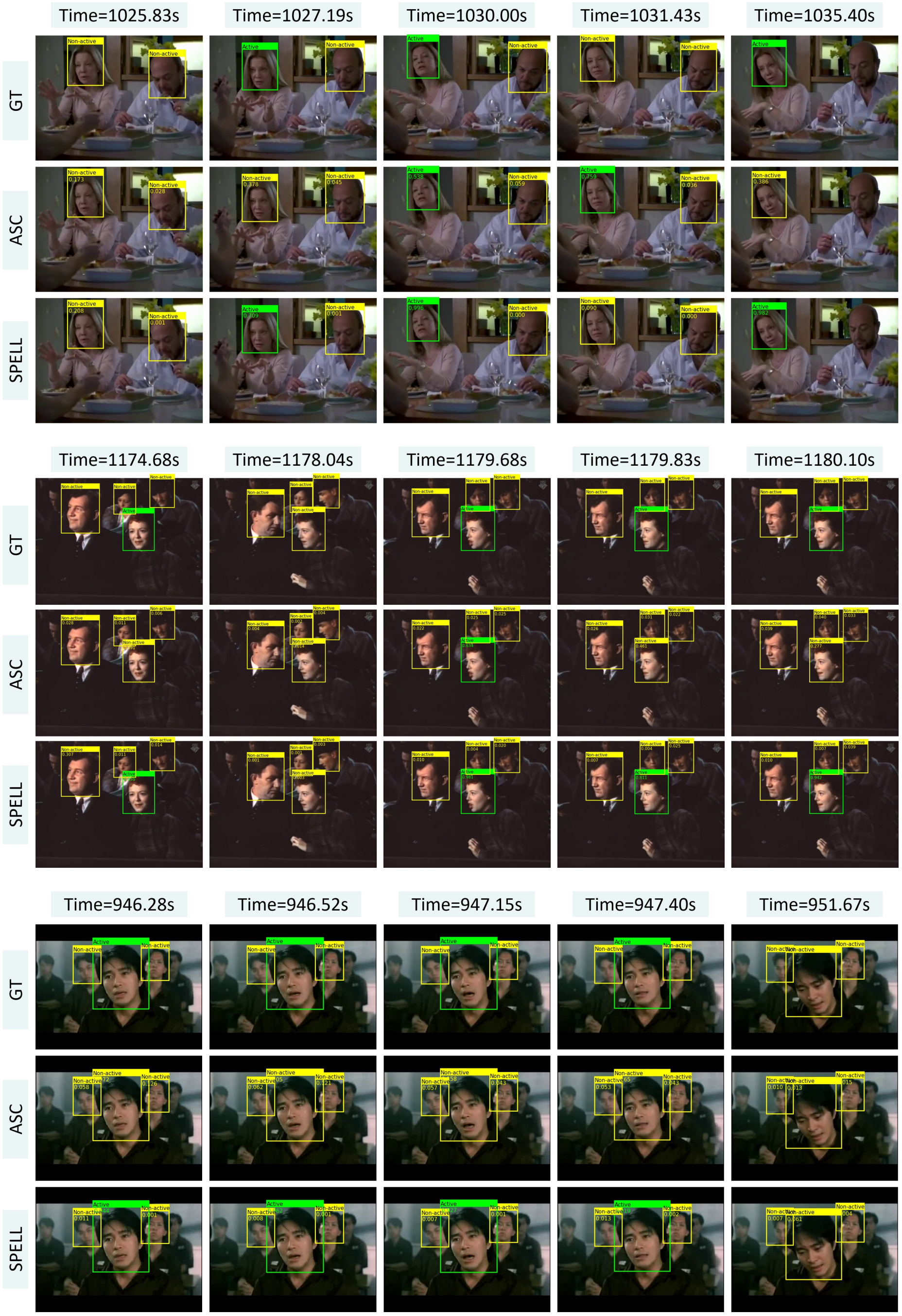

Fig. 1. Qualitative results: The green boxes indicate the active speakers and yellow indicates non-active speaker. The selected frames have a long time-span about 5-10 seconds. GT: Ground Truth.

References

1. Alcazar, J.L., Heilbron, F.C., Mai, L., Perazzi, F., Lee, J.Y., Arbelaez, P., Ghanem, B.: Active Speakers in Context. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. p. 12465–12474 (2020)
2. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
3. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)