JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

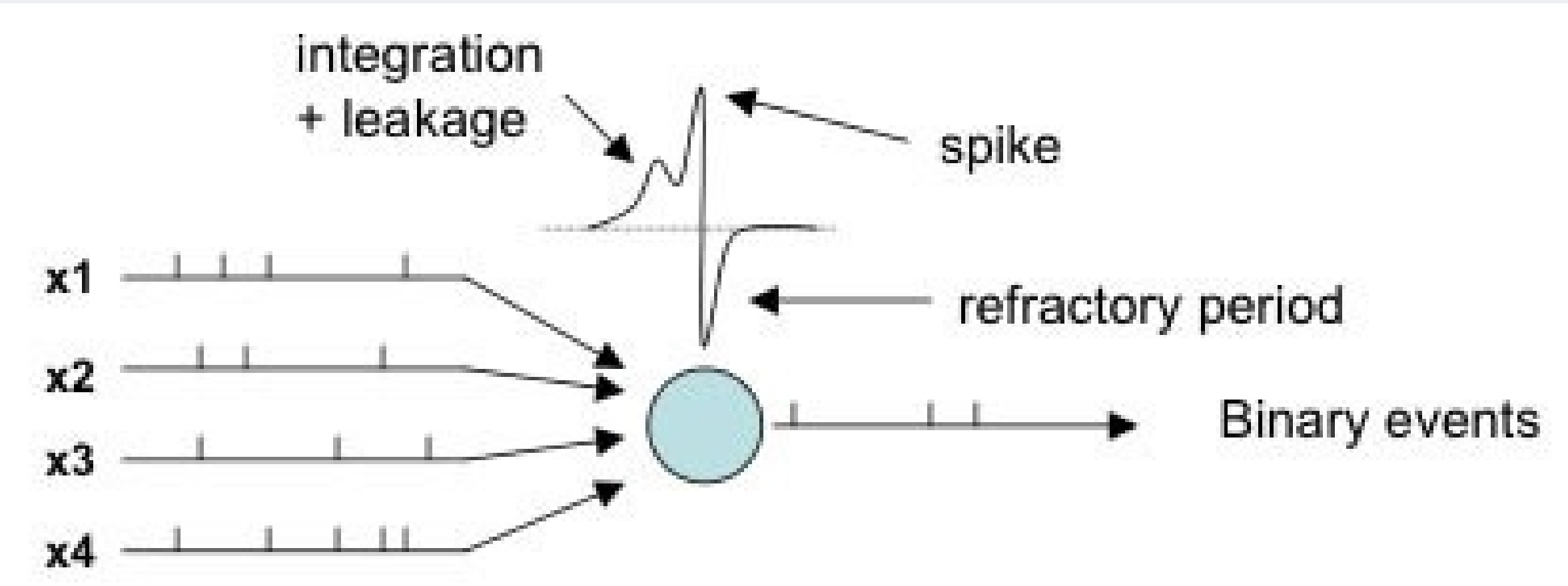
END-TO-END NEUROMORPHIC LIP-READING

Hugo Bulzomi   Marcel Schweiker   Amélie Gruel   Jean Martinet

i3S / CNRS, Université Côte d'Azur, Sophia Antipolis, France

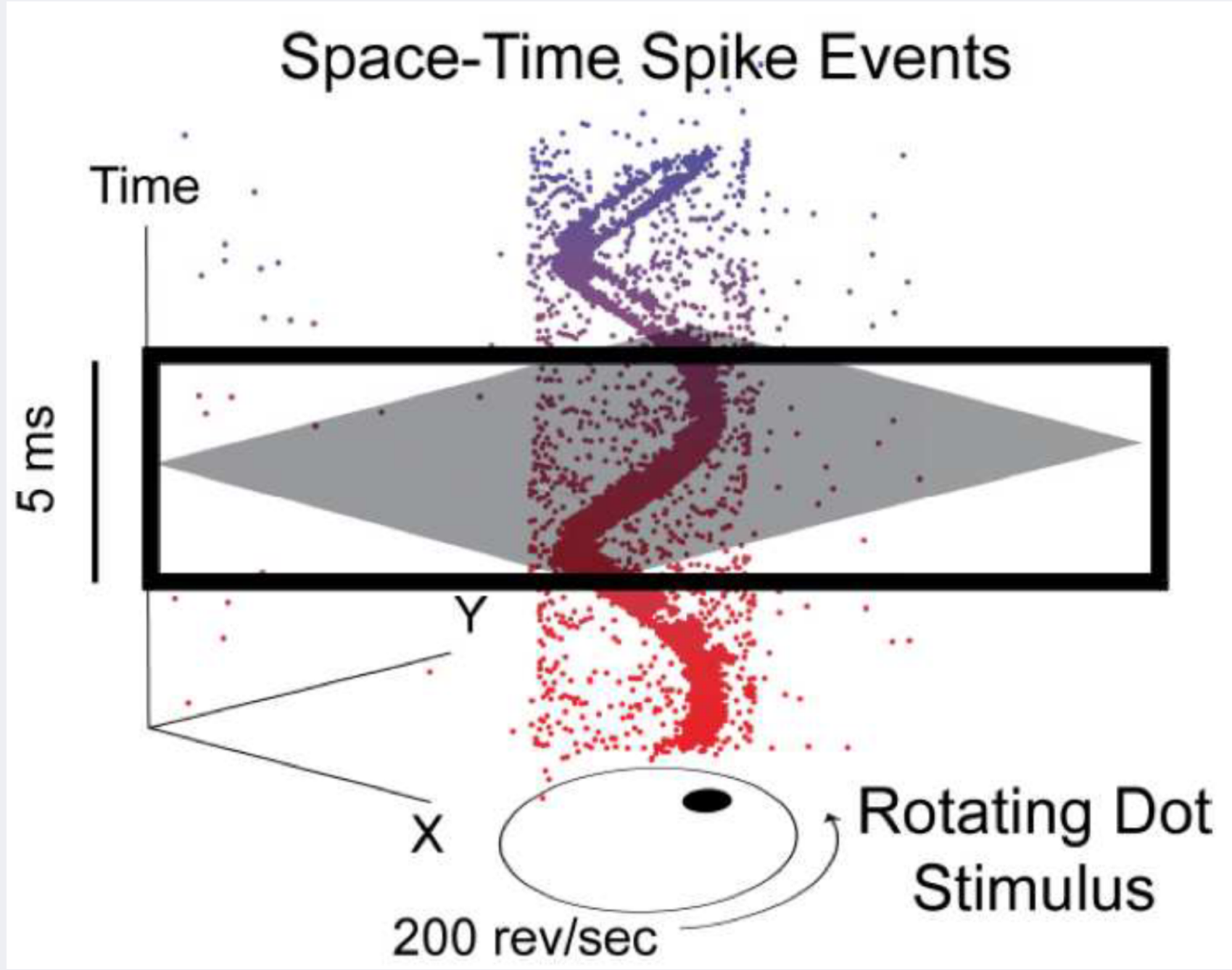
This work was supported by the European Union's ERA-NET CHIST-ERA 2018 research and innovation program under grant agreement ANR-19-CHR3-0008.

Spiking Neural Networks



[Maass, 1997]

Event camera



[Lichtsteiner, 2008] & [Mueggler, 2015]

The authors would like to extend thanks to Yao Lu, Juliette Sabatier, Raphaël Pietrzak, and Quentin Mérilleau for their groundwork on this topic during their internships.

References

[Maass, 1997]: Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. Neural networks

[**Lichtsteiner, 2008]**: Lichtsteiner, P., Posch, C., & Delbruck, T. (2008). A 128x128 120 dB 15μs latency asynchronous temporal contrast vision sensor. IEEE journal of solid-state circuits

[Mueggler, 2015]: Mueggler, E., Gallego, G., & Scaramuzza, D. (2015). Continuous-time trajectory estimation for event-based vision sensors

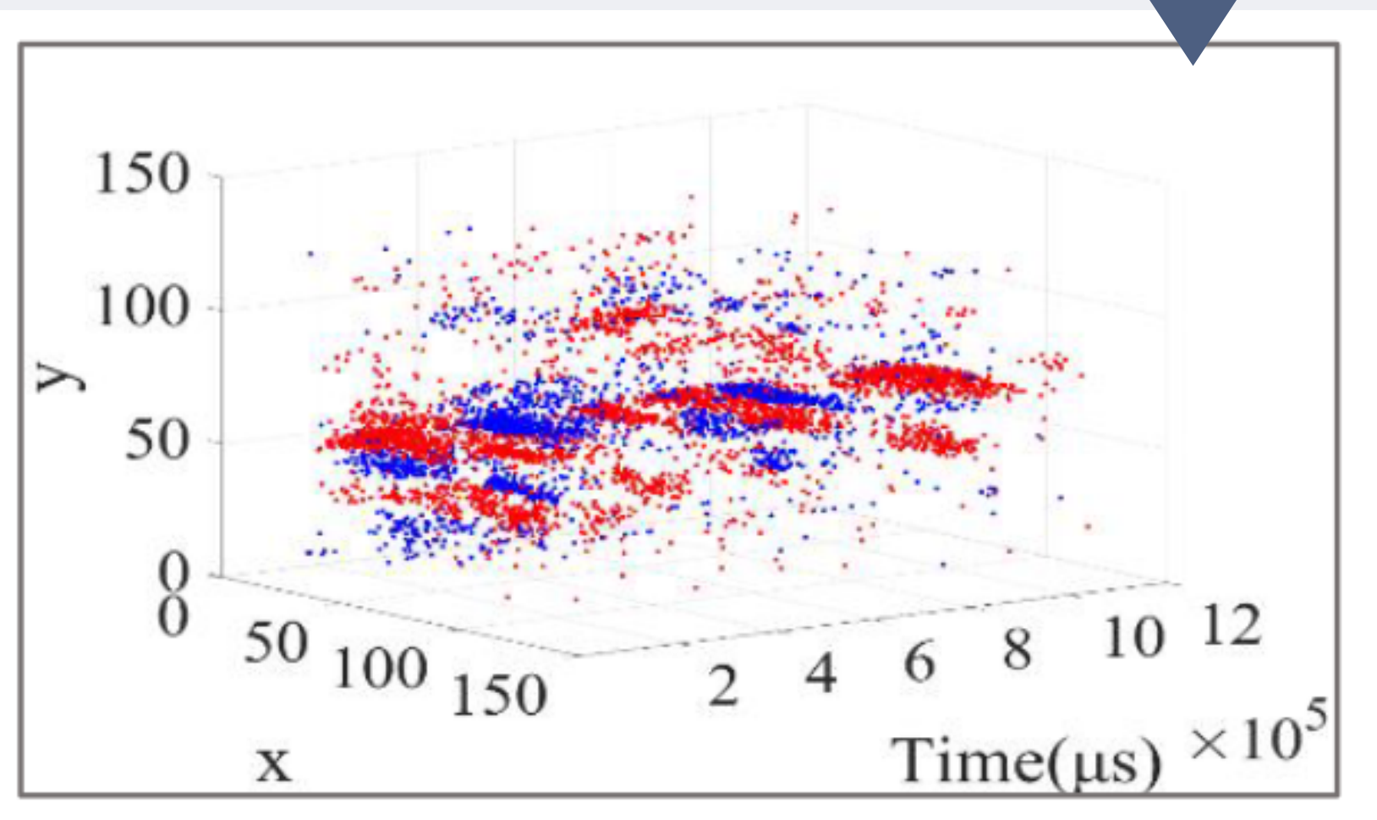
[**Tan et al, 2022]**: Tan, G., Wang, Y., Han, H., Cao, Y., Wu, F., & Zha, Z. J. (2022). Multi-grained spatio-temporal features perceived network for event-based lip-reading. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[Fang et al, 2020]: Fang, H., Shrestha, A., Zhao, Z., & Qiu, Q. (2020). Exploiting neuron and synapse filter dynamics in spatial temporal learning of deep spiking neural network. arXiv preprint arXiv:2003.02944.

PIPELINE

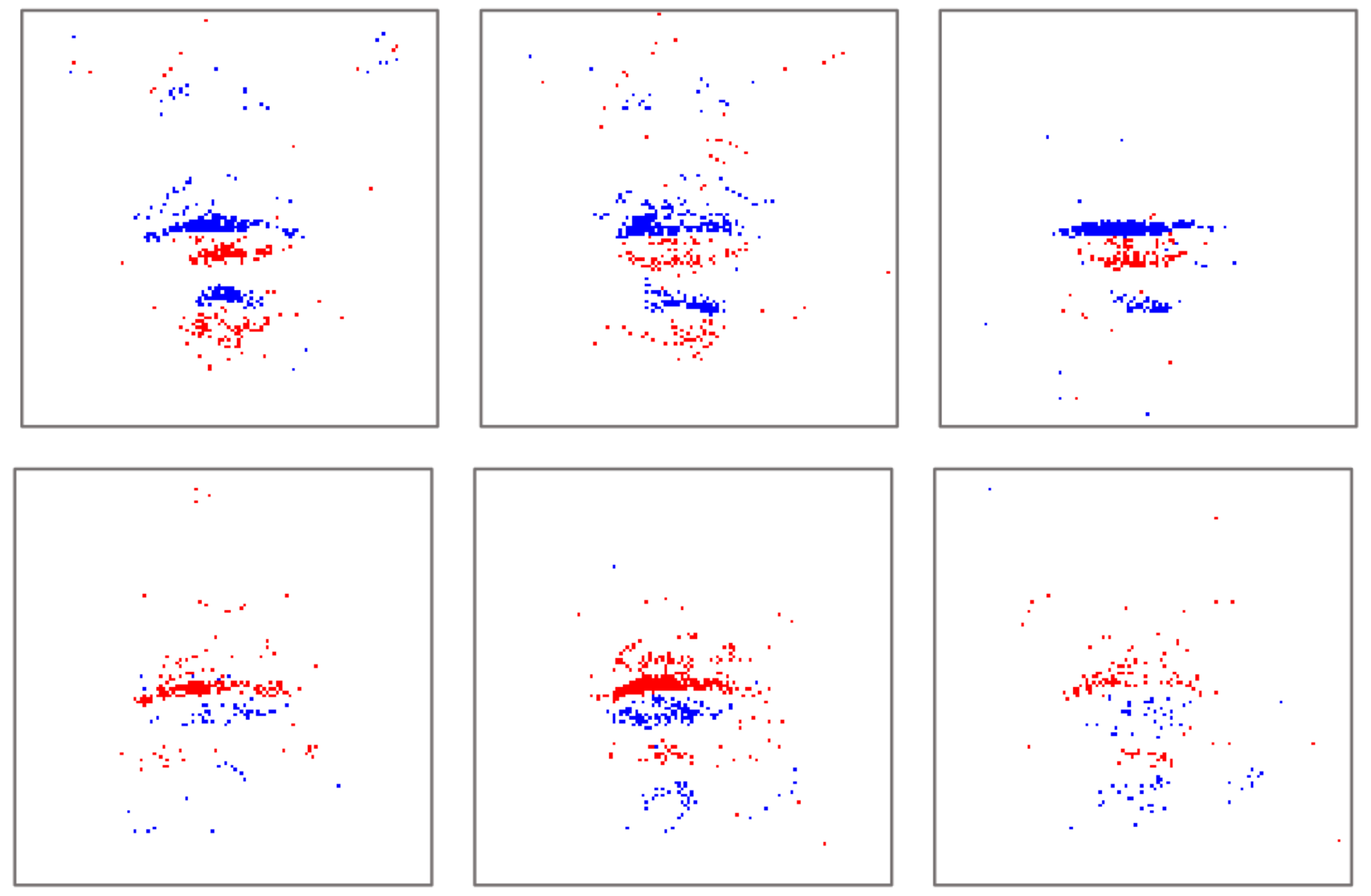
SPEAKER

EVENT CAMERA



(a) Raw Events

Preprocessing to voxel grid



[Tan et al, 2022]

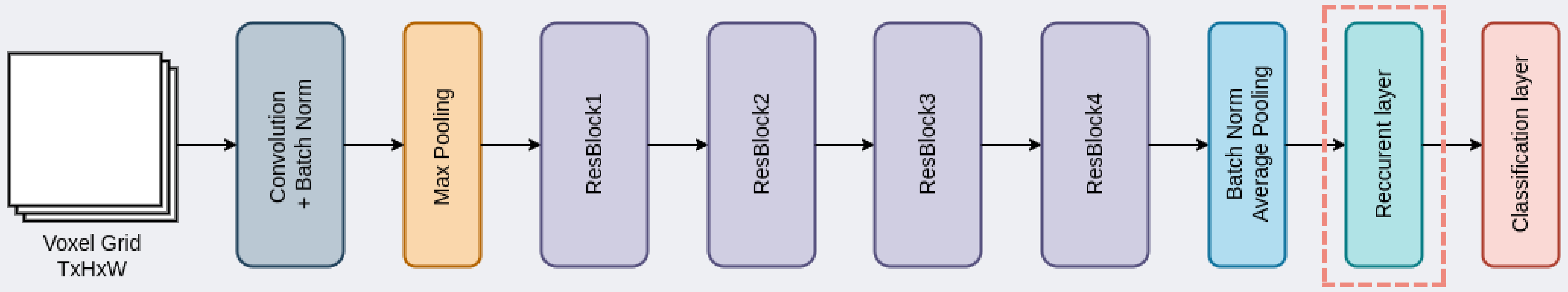
Aggregate events into frames
3D grid (Time, height, width)

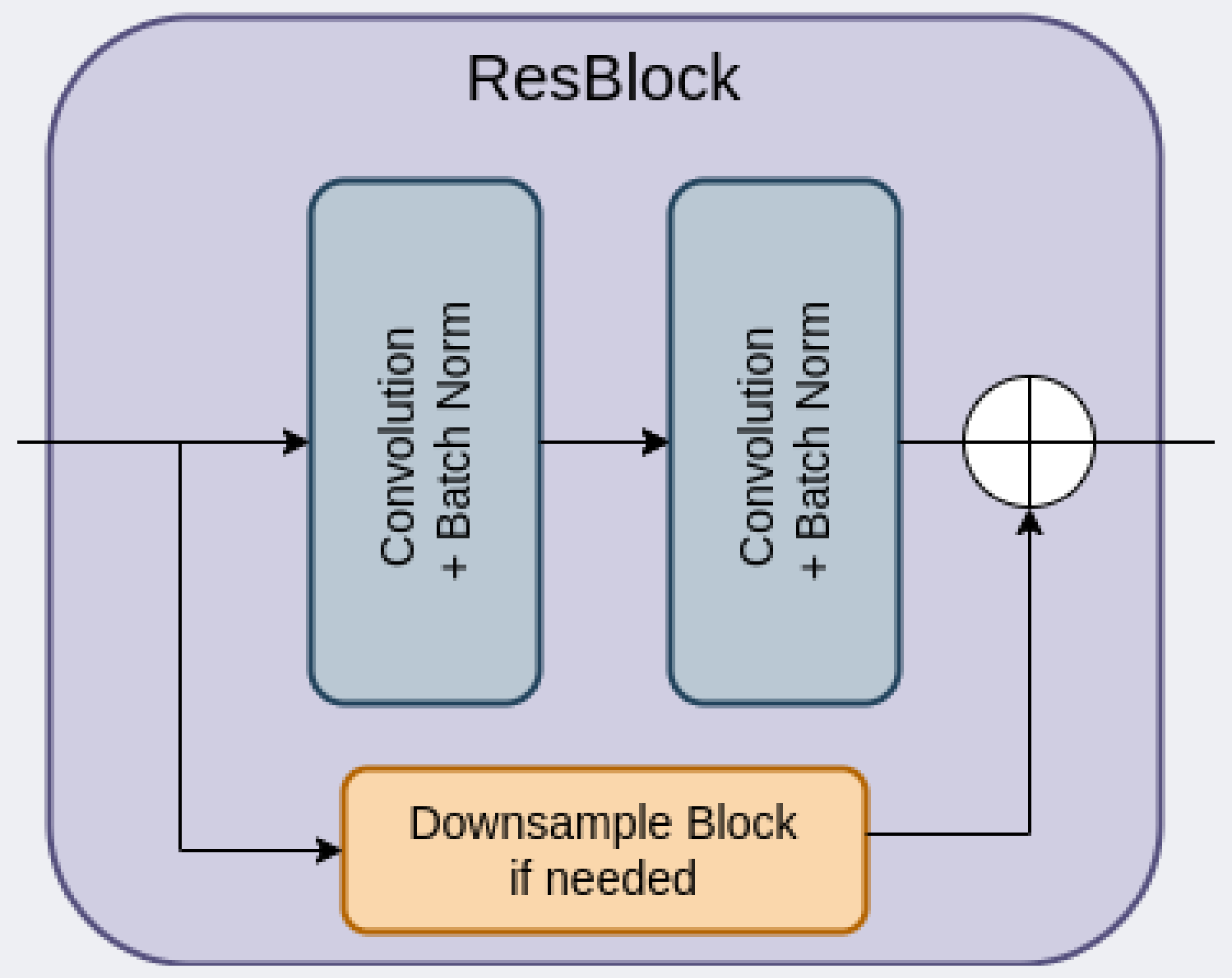
Lip-Reading SNN Model

Prediction

"OYSTER"

PROPOSED LIP-READING SNN

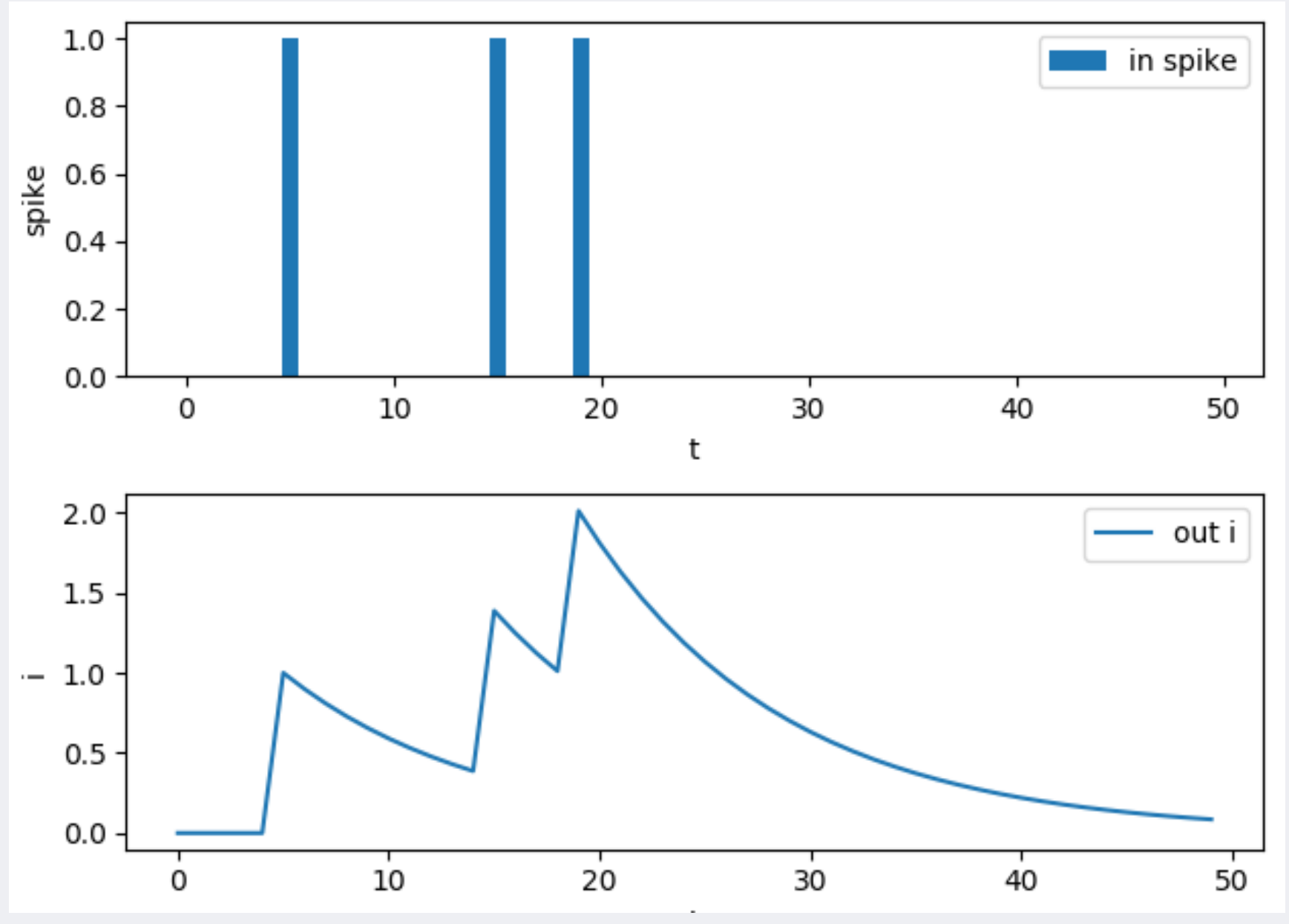




Model	Model	Accuracy
MSTP	**Original ANN [33]**	**0.721**
Spiking MSTP	Simple fc layer	0.522
Spiking MSTP	Spiking bi-GRU	0.463
Spiking MSTP	Spiking recurrent layer	0.476
Spiking MSTP	**Stateful synapse**	**0.602**

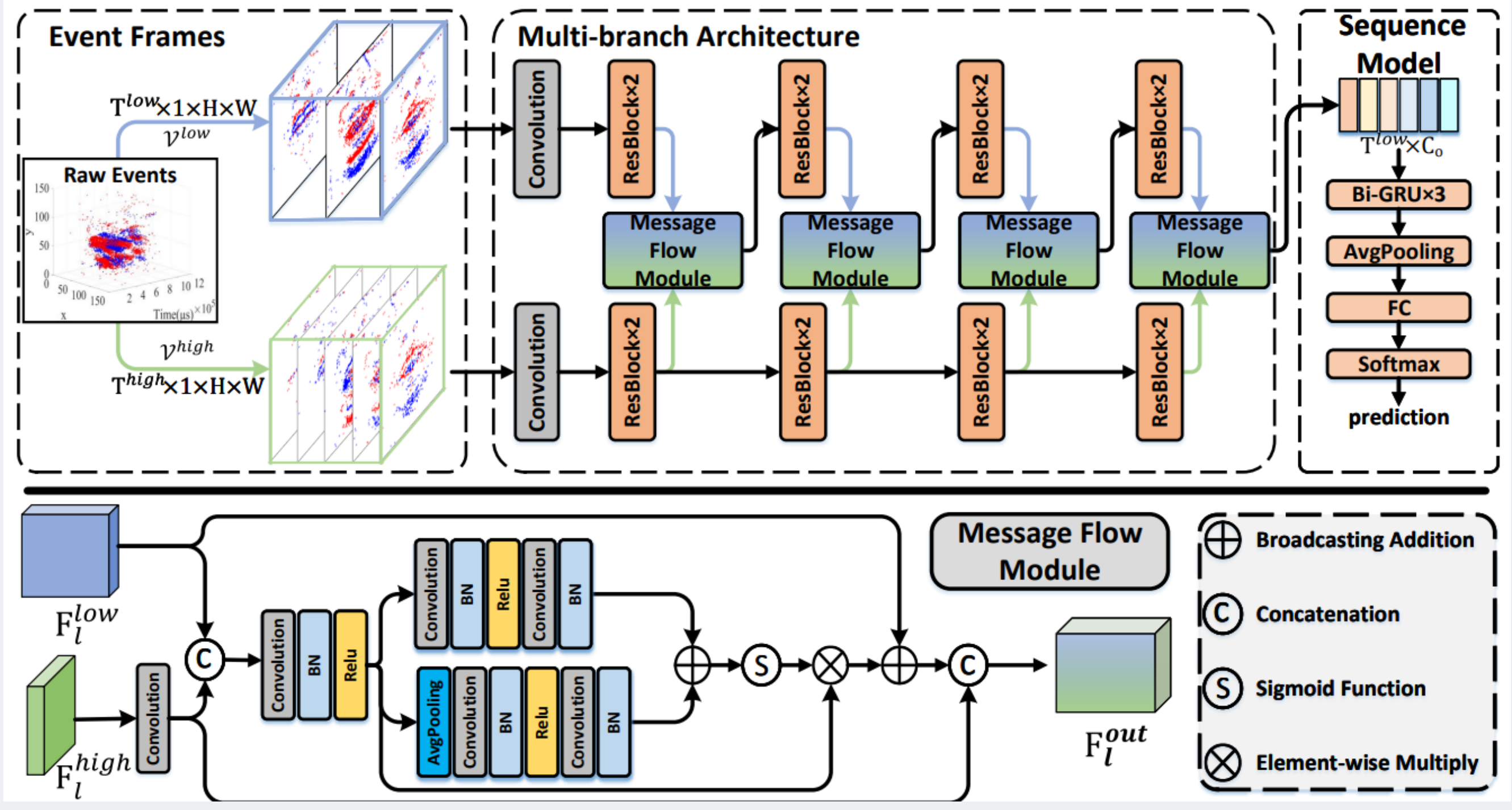
Replacements tested for the original 3-layer GRU

Stateful synapses principle



[Fang et al, 2020]

Inspired by MSTP model



[Tan et al, 2022]

CONCLUSION

We propose the **very first fully neuromorphic** approach for lip-reading using a **spiking neural network.**

Further works might improve this architecture and **investigate other languages**.

Model	Accuracy	Size
MSTP	**0.721**	241.5MB
MSTP low w/out GRU	0.591	**47MB**
SNN1	0.395	**26.7MB**
SNN2	0.514	88.9MB
Spiking MSTP	**0.602**	47MB

Our model needs ~5 times less memory than the SOTA, while retaining 85% of its accuracy.