Spiking neural networks trained with backpropagation for low power neuromorphic implementation of voice activity detection

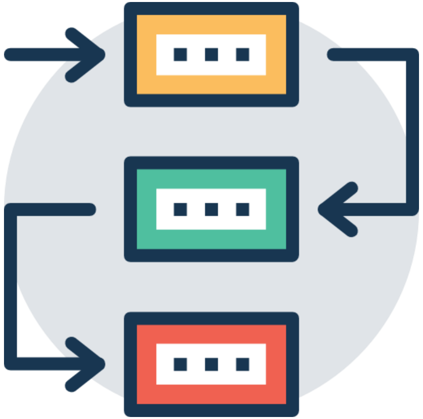
Flavio Martinelli* , Giorgia Dellaferrera*,
Pablo Mainar†, Milos Cernak†

*Ecole Polytechnique Fédérale de Lausanne (EPFL)

†Logitech Europe S.A., Lausanne, Switzerland

ICASSP2020 Barcelona

Outline



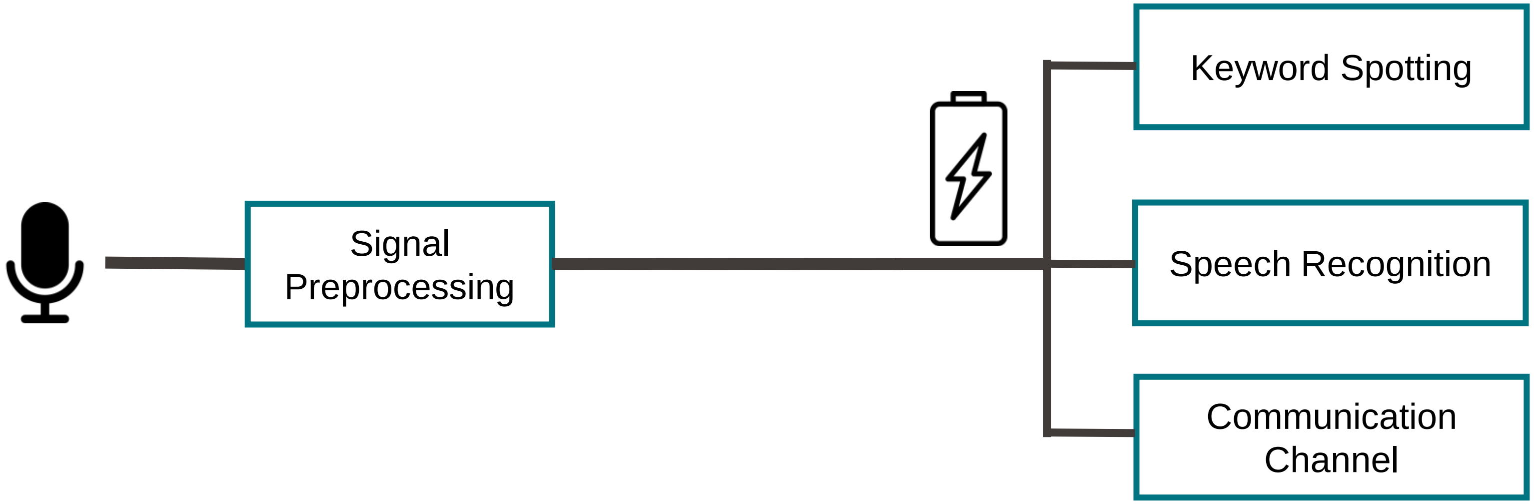
- Introduction to Voice Activity Detection (VAD)
- Current SNN training approaches
- Training SNN with backpropagation
- Pushing the limits of SNNs with temporal coding and lottery tickets
- Results
- Conclusions

Main message

- Neuromorphic microchips are a solution to Voice Activity Detection in battery powered devices

- Spiking network training algorithms should exploit as possible temporal dynamics to achieve lower power consumption

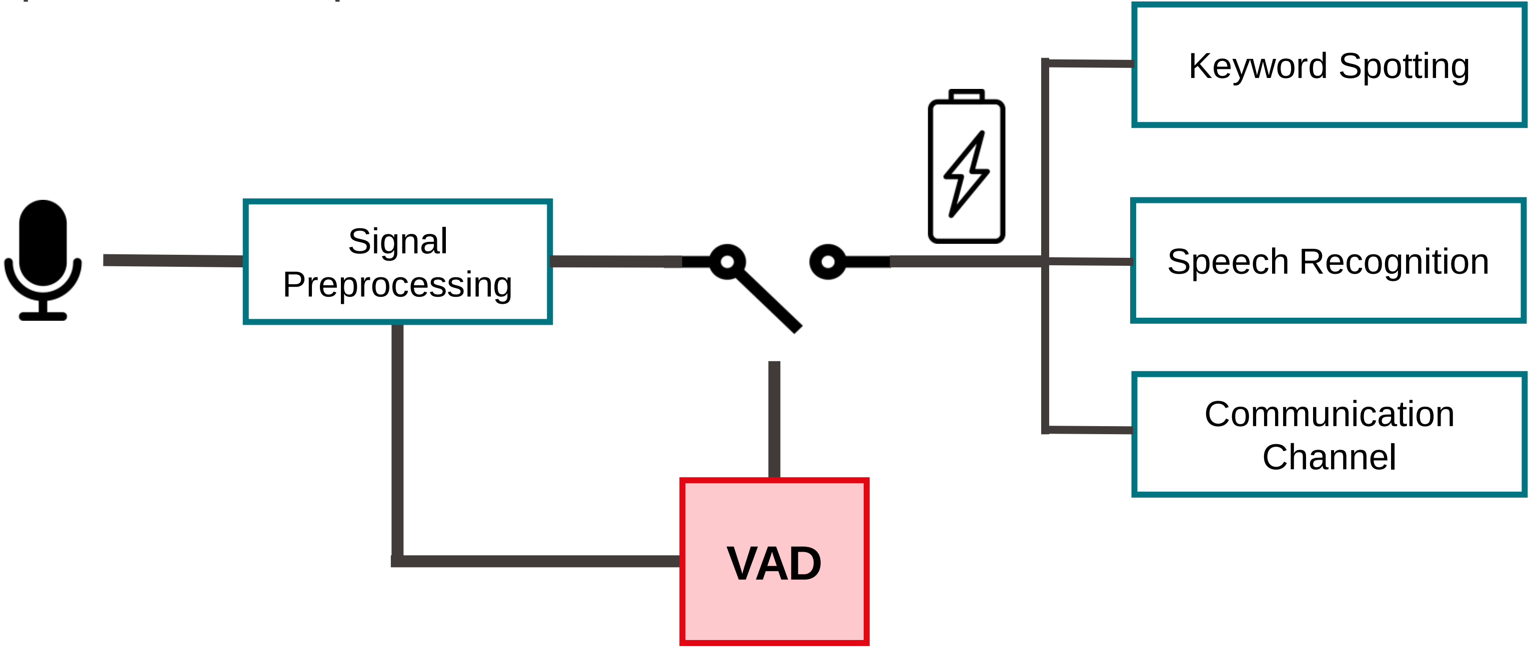
Voice Activity Detection (VAD)



Javier Ramirez et al "Voice activity detection. fundamentals and speech recognition system robustness" 2007

Voice Activity Detection (VAD)

VAD as a **gating system:** limit further computational processing and power consumption



Javier Ramirez et al "Voice activity detection. fundamentals and speech recognition system robustness" 2007

Rate coding vs. temporal coding

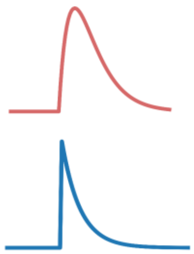
- Most performing training technique: conversion from trained artificial model to spiking model
  - Translation of analog neuron activations into spiking rates
  - Latency – Accuracy tradeoff

- We want to fully exploit the time dimension to encode information
  - Less events to convey the same amount of information
  - Exploit novel computational properties

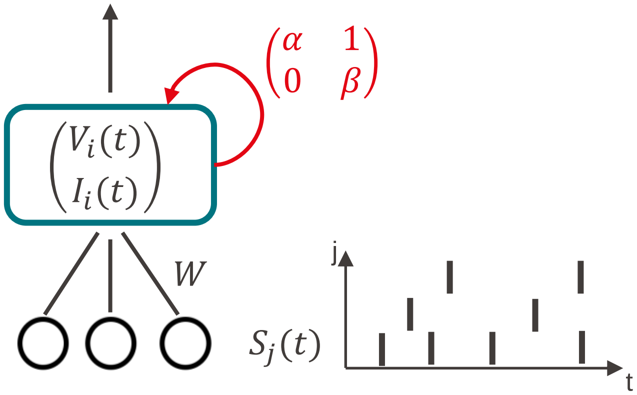
Pfeiffer, Michael, and Thomas Pfeil. "Deep learning with spiking neurons: opportunities and challenges." Frontiers in neuroscience 12 (2018): 774.

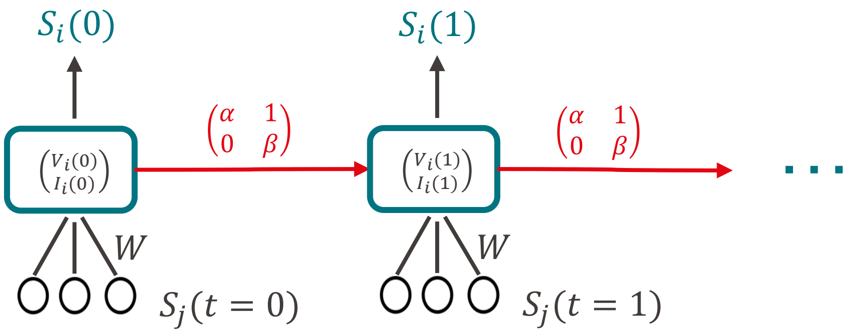
SNN recast as recurrent Network

$$S_i(t) = \Theta(V_i(t) - \theta)$$



$$\underline{V_i(t+\Delta T)} = \alpha V_i(t) + I_i(t) - S_i(t),$$
$$\underline{I_i(t+\Delta T)} = \beta I_i(t) + \sum_j w_{ij} S_j(t)$$

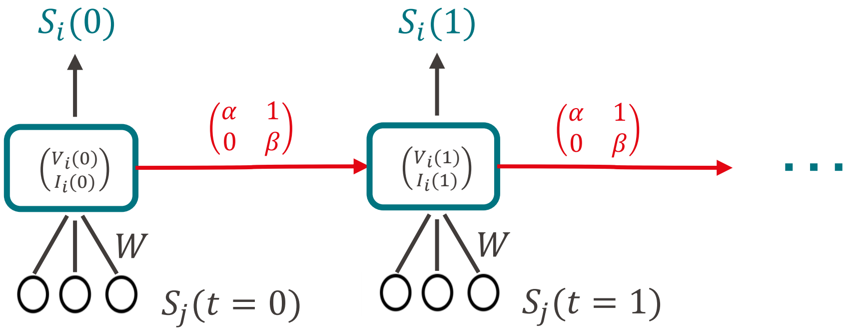




Train with BackPropagation Through Time (BPTT) like a recurrent network

Neftci, Emre O et al. 'Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks'.
Bellec, Guillaume, et al. 'Long short-term memory and learning-to-learn in networks of spiking neurons'.

Surrogate gradients and loss function

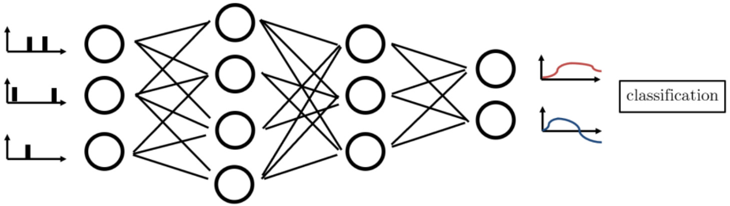


Problem: differentiating through the Heaviside function cancels all the gradients

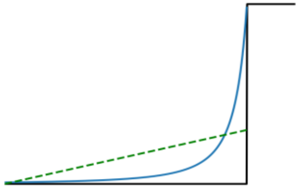
Solution: use a surrogate function for the backward pass

$S_i(t) = \Theta(V_i(t) - \theta)$ Forward pass

$$\frac{1}{(1+\lambda|V_i(t)-\theta|)^2}$$ Backward pass



$$\mathcal{L} = \mathrm{CE}(\max(V_s) - \max(V_n))$$

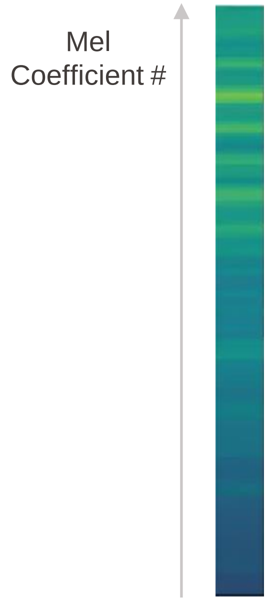


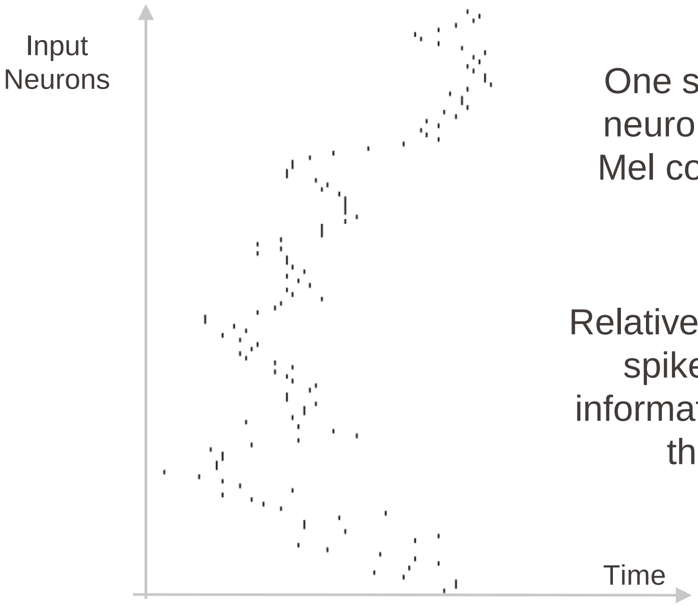
Neftci, Emre O et al. ‘Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks’.
Bellec, Guillaume, et al. ‘Long short-term memory and learning-to-learn in networks of spiking neurons’.

EPFL logitech

Spike temporal encoding of features

64 ms audio frame → 128 log Mel filterbank coefficients → Normalization, discretization and Time To First Spike (TTFS)





One spike per input neuron conveys the Mel coefficient value

Relative timing between spikes is the only information available to the network

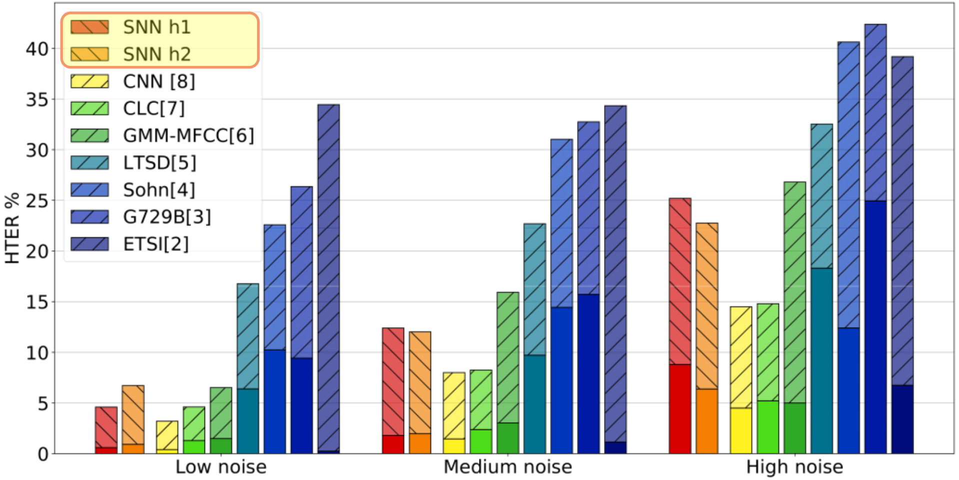
Results on QUT-NOISE-TIMIT

FAR : false alarm rate
MR : miss rate

$$DCF = 0.25\,FAR + 0.75\,MR$$

$$HTER = 0.5\,FAR + 0.5\,MR$$

Method / SNR	+15	+10	+5	0	-5	-10
Sohn [4]	11.1	13.4	19.7	25.9	31.3	37.6
Segbroeck [28]	6.1	6.0	10.4	10.8	18.3	23.2
Neurogram [29]	5.5	5.9	10.2	10.0	17.5	23.7
SNN h1-*w*	2.9	4.5	7.1	9.7	12.5	16.3
SNN h2-*w*	5.0	5.8	7.3	9.6	12.2	15.7
SNN h1	2.4	3.4	5.9	10.2	16.3	26.5
SNN h2	3.9	4.5	6.2	9.4	14.1	21.1



Comparisons:
SNN trained on entire dataset

DCF (Detection Cost Function)

Previous ICASSP work:
Neurogram (ANN based)

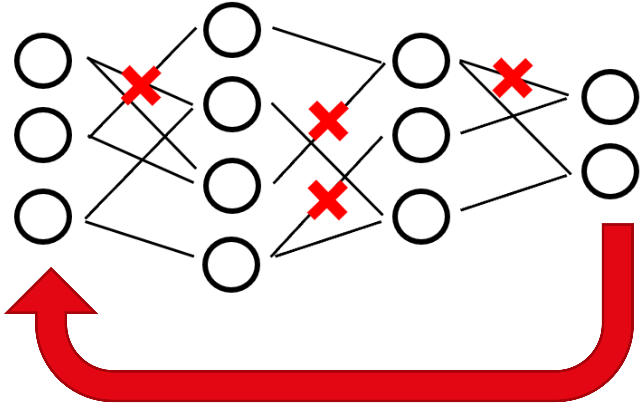
HTER (Half Total Error Rate)

[6-8]: Machine Learning approaches, trained on specific noise level

[2-5]: Standard signal processing solutions

Lottery Ticket Hypothesis and pruning

- **Lottery Ticket Hypothesis**: within the model there exist subnetworks which can achieve the same performance as the full model
  - Lottery ticket subnetworks are defined by the random weight initialization process
  - Ability to not lose performance at **15%** of the original connectivity



	Rate	#Params	Power	HTERs %
SNN h1	1 - 4.7	$26k$	$33.0\,\mu W$	4.6‖12.4‖25.2
SNN h1-p	1 - 2.9	4096	$25.1\,\mu W$	4.7‖12.5‖25.8

Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." arXiv preprint arXiv:1803.03635 (2018).

EPFL

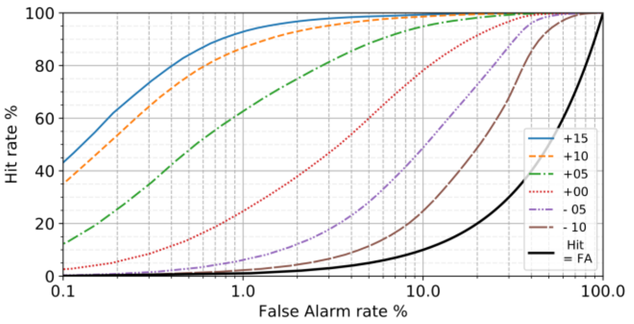
logitech

ROC curve and energy estimation

	Rate	#Params	Power	HTERs %
SNN h1	1 - 4.7	$26k$	$33.0\,\mu W$	4.6‖12.4‖25.2
SNN h1-p	1 - 2.9	4096	$25.1\,\mu W$	4.7‖12.5‖25.8

- Low average spiking rate
- Estimated power consumption in the order of tens of uW (computed from TrueNorth power consumption profile and adapted to our network size and activity)
- Pruning effective in reducing network average activity and Synaptic Operations
- Low power VAD implementations reach up to less than 1uW but at the performance cost of performance: (84% hit rate and 72% correct rejection @5dB against our SNN which respectively has 97% and 84%)

$$\max(V_{speech}) - \max(V_{no\_speech}) > \rho$$



Paul Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science, vol. 345, no. 6197, pp. 668–673, 2014.

Minchang Cho et al., "17.2 a 142nw voice and acoustic activity detection chip for mm-scale sensor nodes using time interleaved mixer-based frequency scanning,"

Conclusions

- We proposed a VAD spiking model that competes with state of the art, with a good tradeoff between power consumption and performance. SNN VAD solution from * scores at 26mW of power consumption

- We pushed temporal coding to the limit and showed that spiking networks can work with complex real valued features coded in the temporal domain

- Addressed neuromorphic chips connectivity problems with pruning techniques

** Steven K Esser, Paul A Merolla, John V Arthur, Andrew S Cassidy, et al., “Convolutional networks for fast energy-efficient neuromorphic computing,” Proc. Nat. Acad. Sci. USA, vol. 113, no. 41, pp. 11441–11446, 2016.*

Limitations and Thought for the future

- Training process is computationally very expensive due to the amount of timesteps to backpropagate
- Translation into neuromorphic implementations difficult due to different specifications of each manufacturer

- Use of more elaborate spike encodings and loss functions
- Exploit recurrent connections
- Test on harder tasks such as keyword spotting

Bibliography and Attributions

Javier Ramirez et al "Voice activity detection. fundamentals and speech recognition system robustness" 2007

Pfeiffer, Michael, and Thomas Pfeil. "Deep learning with spiking neurons: opportunities and challenges." Frontiers in neuroscience 12 (2018): 774.

Neftci, Emre O et al. 'Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks'.

Bellec, Guillaume, et al. 'Long short-term memory and learning-to-learn in networks of spiking neurons'.

Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." arXiv preprint arXiv:1803.03635 (2018).

Paul Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science, vol. 345, no. 6197, pp. 668–673, 2014.

Minchang Cho et al., "17.2 a 142nw voice and acoustic activity detection chip for mm-scale sensor nodes using time interleaved mixer-based frequency scanning,"

** Steven K Esser, Paul A Merolla, John V Arthur, Andrew S Cassidy, et al., "Convolutional networks for fast energy-efficient neuromorphic computing," Proc. Nat. Acad. Sci. USA, vol. 113, no. 41, pp. 11441–11446, 2016.*

- Slides [3]: icons made by Prosymbols from www.flaticon.com
- Slides [4-5]: icons made by Kiranshastry, Freepik, prettycons from www.flaticon.com

Acknowledgements

Giorgia Dellaferrera

PreDoctoral Researcher (PhD) at IBM Research Zurich

Pablo Mainar

Logitech Europe S.A.

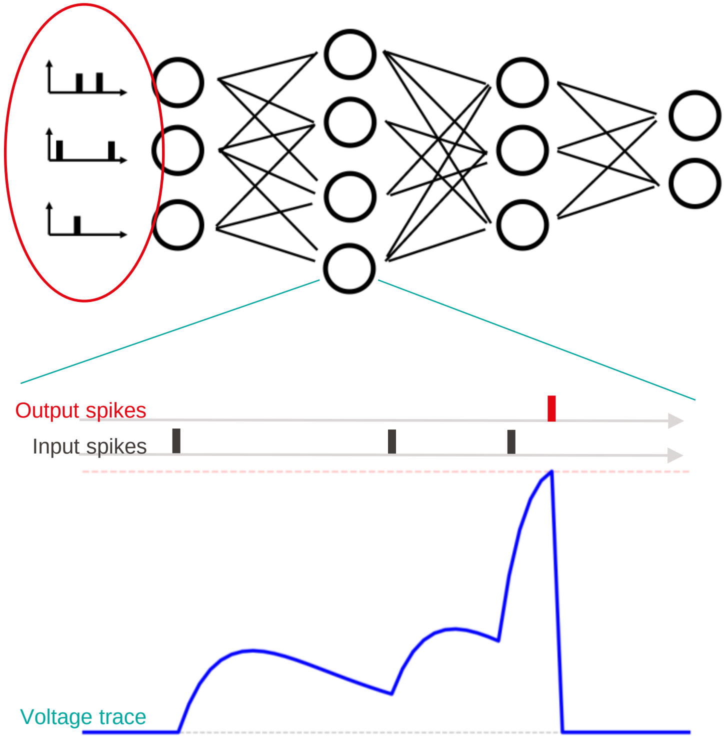
Milos Cernak

Logitech Europe S.A.

Extra slides that did not fit the 15 minutes

Spiking Neural Networks (SNNs)

- Spiking neurons have **time dynamics**
- Binary activation function called **spike**
- Asynchronous, **event driven** processing
- Leaky integrator dynamics
- More biologically realistic than artificial networks



Why spiking neurons for VAD in battery powered devices?

- VAD is an always on system
- By design it needs to be power efficient and have a good performance
- Embedded neuromorphic microchips are a low power and efficient solution

FEATURE ARTICLE: Neuromorphic Computing

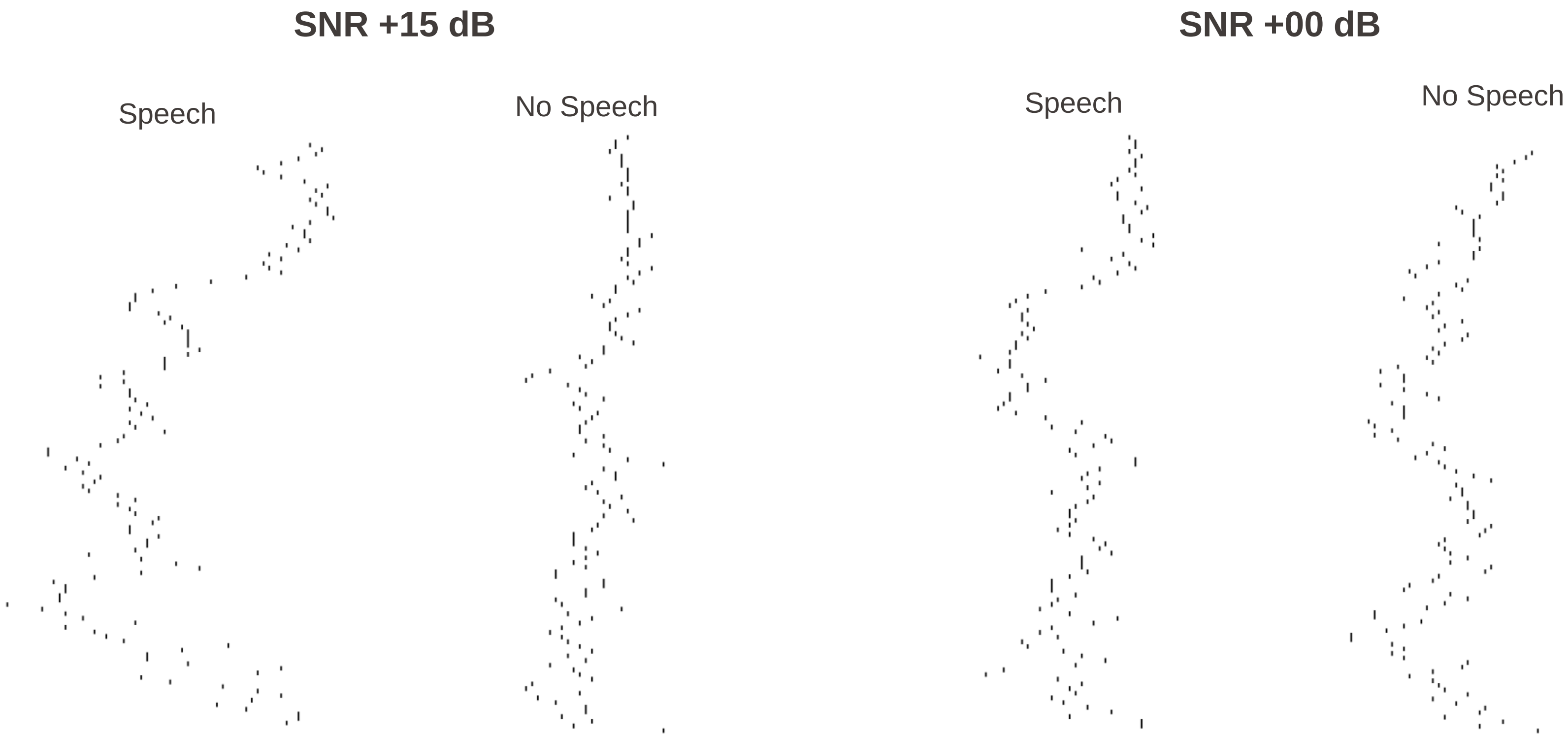
Loihi: A Neurom
Manycore Proce
On-Chip Learnin

ARTIFICIAL BRAINS

A million spiking-neuron integrated circuit with a scalable communication network and interface

Paul A. Merolla,[1*] John V. Arthur,[1*] Rodrigo Alvarez-Icaza,[1*] Andrew S. Cassidy,[1*] Jun Sawada,[2*] Filipp Akopyan,[1*] Bryan L. Jackson,[1*] Nabil Imam,[3] Chen Guo,[4] Yutaka Nakamura,[5] Bernard Brezzo,[6] Ivan Vo,[2] Steven K. Esser,[1] Rathinakumar Appuswamy,[1] Brian Taba,[1] Arnon Amir,[1] Myron D. Flickner,[1] William P. Risk,[1] Rajit Manohar,[7] Dharmendra S. Modha[1]†

Spike pattern examples



Network models and classification

Network	Architecture	τ_I	τ_V	Classification
SNN h1	128 – 200 – 2	5	10	Frame by frame. Median smoothing on 11 predictions
SNN h2	128 – 100 – 15 - 2	5	10 - 300	5 successive frames, classification on the last one. Median smoothing on 11 predictions