Micro-power spoken keyword spotting on Xylo Audio 2

Thurgauerstrasse 40, 8050 Zürich, Switzerland
Hannah Bos, Dylan Muir

*We describe the implementation of a spoken audio keyword-spotting (KWS) benchmark "Aloha" on the Xylo Audio 2 (SYNS61210) Neuromorphic processor device. We obtained high deployed quantized task accuracy, (95%), exceeding the benchmark task accuracy. We measured real continuous power of the deployed application on Xylo. We obtained best-in-class dynamic inference power (*291 µW*) and best-in-class inference efficiency (*6.6 µJ/Inf*). Xylo sets a new minimum power for the Aloha KWS benchmark, and highlights the extreme energy efficiency achievable with Neuromorphic processor designs.*

Contents

Audio processing with Xylo™

Xylo™ Audio is a family of ultra-low-power audio inference chips, designed for in- and near-microphone analysis of audio in real-time energy-constrained scenarios. Xylo is designed around a highly efficient integer-logic processor which simulates parameter- and activity-sparse spiking neural networks (SNNs) using a leaky integrate-and-fire (LIF) neuron model. Neurons on Xylo are quantised integer devices operating in synchronous digital CMOS, with neuron and synapse state quantised to 16 bit, and weight parameters quantised to 8 bit. Xylo is tailored for real-time streaming operation, as opposed to accelerated-time operation in the case of an inference accelerator. Xylo Audio includes a low-power audio encoding interface for direct connection to a microphone, designed for sparse encoding of incident audio for further processing by the inference core.

In this report we present the results of a spoken KWS audio benchmark deployed to Xylo Audio 2. We describe the benchmark dataset; the audio preprocessing approach; and the network architecture and training approach. We present the performance of the trained models, and the results of power and latency measurements performed on the Xylo Audio 2 development kit. We include for comparison previous benchmarks of the Aloha KWS task on other neuromorphic devices, mobile inference processors, CPUs and GPUs.

The Aloha Benchmark

We implemented the "Aloha" benchmark dataset introduced by Blouw et. al [1]. This includes a training set of approximately 2000 utterances from 96 speakers, with a 3:1 ratio between the target phrase ("aloha") and non-target phrases. Figure 1 shows the distribution of sample durations in the training and test datasets. In the results below we use the provided test set of 192 samples.

We designed a network that identified "aloha" target samples by producing one or more output events, and non-target samples by remaining silent. We clipped or extended samples to 3 s by padding with silence.

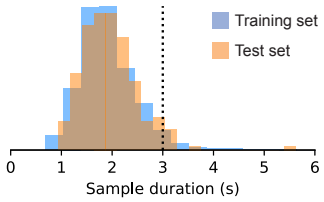

Figure 1: *Distribution of sample durations in the Aloha dataset. In this work we pad and clip samples to a uniform* 3 s *duration (dashed line). This retains the majority of data in both train and test datasets.*

Audio preprocessing

We encoded each sample as sparse events, using a simulation of the audio encoding hardware present on the Xylo Audio device. The design of this preprocessing block is shown in Figure 2 [5]. Briefly, this block is a streaming-mode buffer-free encoder, designed to operate continuously on incoming audio. A low-noise amplifier with a selectable gain of 0, 6 or 12 dB amplifies the incoming audio. A band-pass filter bank with 2nd-order Butterworth filters splits the signal into 16 bands, with centre frequencies spanning 40–16 940 Hz and with a Q of 4. The output of these filters is rectified, then passed through a

leaky integrate-and-fire (LIF) neuron to smooth the signal and convert it to events. The result is to convert a single audio channel into 16 sparse event channels with event rate in each channel corresponding to the energy in each frequency band.

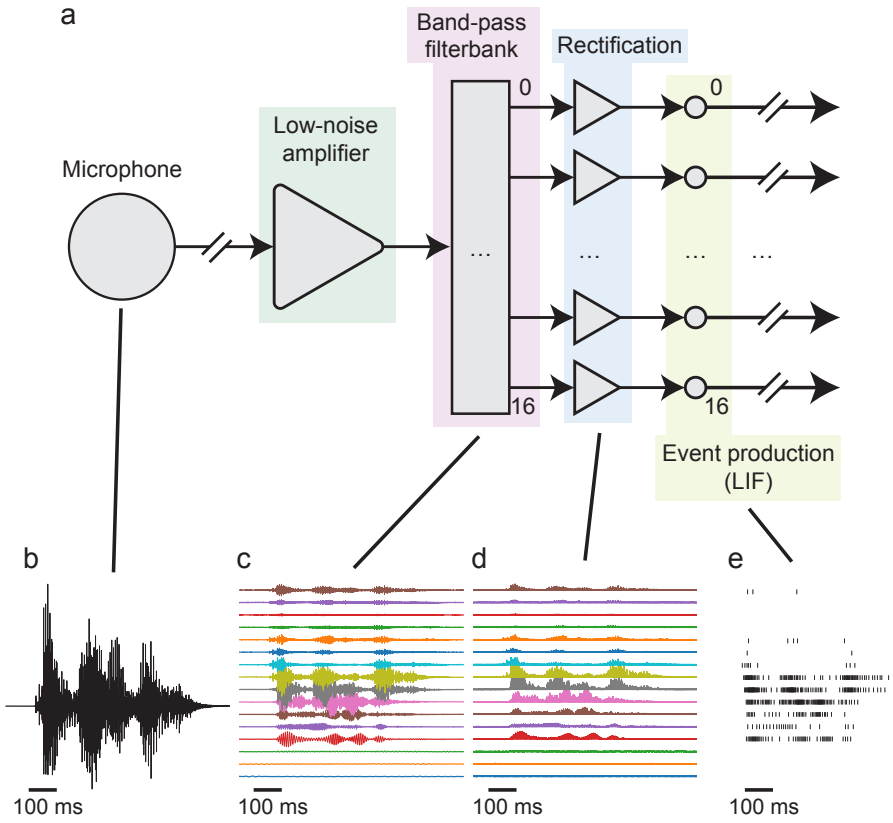

Figure 2: *Audio preprocessing approach. a The stages of audio preprocessing in Xylo Audio 2. Single-channel audio arrives at a microphone (b). This passes through a band-pass Butterworth filterbank, and is split into $N = 16$ frequency bands (c). Filter output is rectified (d) before passing through a bank of LIF neurons that smooth and quantize the signals in each band. The result is a set of sparse event channels (e), where the firing intensity in each channel is proportional to the instantaneous energy in each frequency band.*

Samples were trimmed to 3 s, encoded using the preprocessing block described here, and binned temporally to 100 ms. Our approach operates in streaming-mode, analysing a continuous time-frequency representation of the input audio, similar to a real-time Fourier transform (see Figure 2e).

Network architecture

We use a feed-forward spiking neural network architecture called "SynNet" [2] (Figure 3). This is a fully-connected multi-layer architecture, interleaving linear weight matrices with LIF neuron layers. Each layer has a number of synaptic time constants, where the time constants are defined as $\tau_n = 2^n * 10\,\text{ms}$, and neurons are evenly distributed with the range of time constants for that layer. Layers with two time constants therefore have half the neurons with synaptic time constants $\tau_1 = 20\,\text{ms}$ and half with $\tau_2 = 40\,\text{ms}$. Layers with 4 time constants have a quarter of the neurons with synaptic time constants $\tau_1 = 20\,\text{ms}$; a quarter with $\tau_2 = 40\,\text{ms}$; a quarter with $\tau_3 = 80\,\text{ms}$ and so on. Membrane time constants for all neurons, as well as readout neuron time constants, are set as $\tau_m = 20\,\text{ms}$.

We describe a given SynNet network architecture in the following by defining the list of hidden layer widths H and corresponding list of numbers of time constants τ. For example, the network $H = [160, 60, 60, 60, 60, 60]$ $\tau = [2, 2, 4, 4, 8, 8]$ has 6 hidden layers with a first hidden layer width of 160 neurons, followed by 60 neurons, and so on; and with the first hidden layer containing 2 synaptic time constants, the second with 2 synaptic time constants, the third with 4 and so on. One readout LIF neuron is present in each network, designed to be active for the target class (the keyword "aloha") and inactive for any non-target audio.

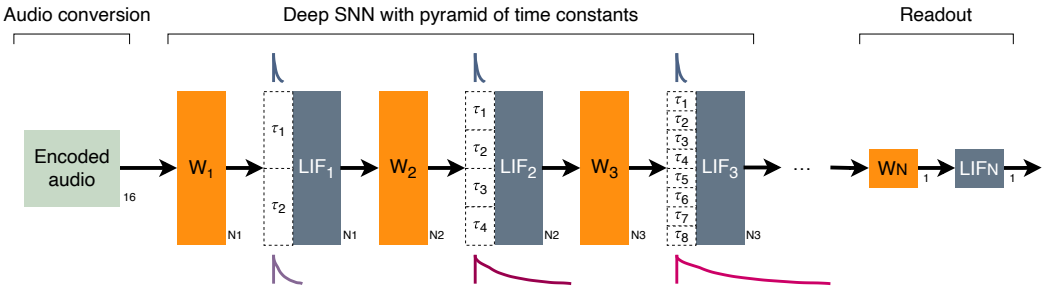

Figure 3: *The SynNet architecture used in this benchmark. Event-encoded audio is provided as input, as described in Figure 2. The network consists of a single feed-forward chain of fully-connected layers, using the LIF neuron model. Several time constants are distributed over each layer, with shorter time constants in early layers and longer time constants in later layers (see text for details). A single readout LIF neuron is used in each network.*

Training

Networks were defined using the open-source Rockpool toolchain (`https://rockpool.ai`), with the *torch* back-end. During training, the membrane potential of the readout neuron was taken as the network output. Targets were defined as $y = 0$ for a non-target sample and $y = 1$ for a target sample.

The training loss for readout channels was defined as follows.

$$\text{PeakLoss}(\mathbf{x}, y) = \begin{cases} \text{MSE}\left(1/M \int_m^{m+M} \mathbf{x}, \mathbf{g}\right) & \text{if } y = 1 \\ w_l \cdot \text{MSE}\left(\mathbf{x}, 0\right) & \text{if } y = 0 \end{cases}$$

where $\mathbf{x}$ is a membrane potential vector over time for a single readout channel; y is the target for the channel (either 1 indicating a target for this channel in this sample, or 0 indicating a non-target for this sample); MSE is the mean-squared-error loss function; $m = \arg\max \mathbf{x}$ is the index of the peak value in $\mathbf{x}$; M is the window duration to examine from $\mathbf{x}$ following the peak; $\mathbf{g}$ is a vector $g \cdot 1$, a target value that $\mathbf{x}$ should match around its peak; w_l is a weighting for the non-target loss component; 0 is the vector of all-zeros. For the networks trained here, we took $M = 140\,\text{ms}$, $g = 1.5$ and $w_l = 1.4$.

Models were trained for 300 epochs, using the PyTorch Lightning package to manage training.

Model performance

Several trained models with a range of total model sizes were evaluated for task performance on the test set. The presence of any readout events during a sample was taken as a "target" prediction. We computed the true positive and false positive rates over the test set, as well as the binary accuracy. The model performance for several model sizes is shown in Table 1.

We computed ROC (Receiver Operator Characteristic) curves for the trained models on the test set, by varying the threshold of the output neuron (Figure 4).

Power and inference rate

Models were quantized and deployed to Xylo devices using the Rockpool deployment pipline. Power was measured on the benchmark application deployed to a Xylo Audio 2 device, on the Xylo Audio 2 hardware development kit (Figure 5), during streaming continuous analysis of the Aloha test-set. The master clock frequency for Xylo Audio was set to 6.25 MHz. Current measurements were taken using on-board current monitors, at a frequency of 1280 Hz. Active power was measured while streaming encoded audio for the entire test set to the Xylo device. Idle power was measured by deploying a model to the device, then measuring consumed power for five seconds with no model input. Inference rate was defined in line with Blouw et al., with one inference corresponding to the processing of 10 time-steps by the network [1].

For all trained models, Xylo required idle power of 216–217 µW and active power of 468–514 µW, resulting in dynamic power of 251–298 µW. Inference rate varied with model size, ranging 40–102 Inf/s. This corresponds to dynamic energy per inference of 2.4–7.3 µJ/Inf. Considering active energy per inference, we computed a range of 4.6–12.7 µJ/Inf.

Comparison with other inference devices

We compared our results with several other hardware deployments of the same benchmark task (Table 2; Figure 6). The model deployed to Xylo Audio exhibited the lowest continuous idle, active and dynamic power consumption of any of the comparison devices. Previous results for the Aloha benchmark report dynamic energy per inference; Xylo Audio achieved the lowest dynamic energy per inference of all comparison devices.

For the devices where total active power was reported, we also compared active energy required per inference, as we believe this is a more realistic system-level metric. Xylo Audio achieved the lowest active energy per inference by an order of magnitude (Table 2 Act. E).

N_{tot}	Model	Acc.	TPR	FPR
461	$H = [160, 60, 60, 60, 60, 60]$	96.88%	97.92%	4.17%
411	$H = [110, 60, 60, 60, 60, 60]$	97.92%	97.92%	2.08%
401	$H = [150, 50, 50, 50, 50, 50]$	97.40%	93.75%	6.25%
361	$H = [60, 60, 60, 60, 60, 60]$	94.27%	97.92%	8.33%
341	$H = [140, 40, 40, 40, 40, 40]$	98.44%	100.0%	3.12%
281	$H = [130, 30, 30, 30, 30, 30]$	93.23%	94.79%	8.33%
221	$H = [120, 20, 20, 20, 20, 20]$	97.40%	95.83%	1.04%
$^{\dagger}461$	$H = [160, 60, 60, 60, 60, 60]$	95.31%	91.67%	1.04%

Table 1: *Test performance for trained models. Time constants* $\tau = [2, 2, 4, 4, 8, 8]$ *for all models.* N_{tot}*: Total neurons in the model; Acc.: Accuracy; TPR: True Positive Rate* $TPR = TP/(TP + FN)$*; FPR: False Positive Rate* $FPR = FP/(FP + TN)$*.* †*Quantised model result deployed to the Xylo architecture.*

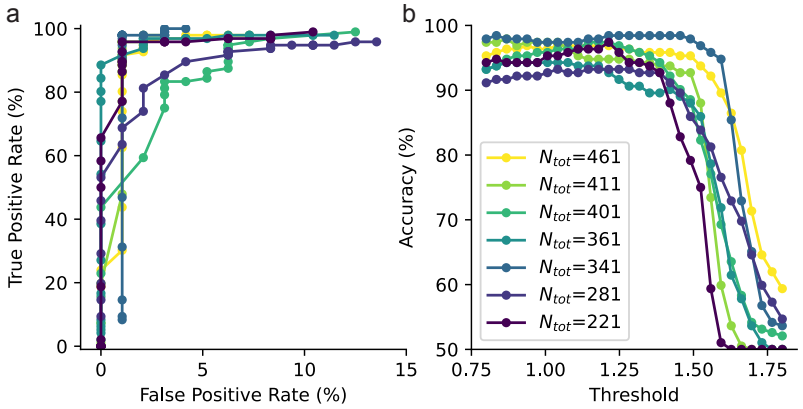

Figure 4: *ROC curves for the trained models in Table 1. a True Postive Rate vs False Positve Rate curves. b Accuracy for the several models while varying the threshold of the readout neuron.*

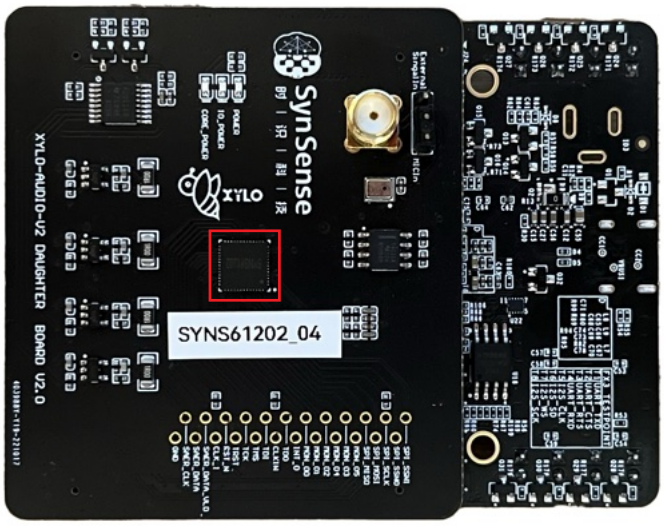

Figure 5: *The Xylo™ Audio 2 hardware development kit (HDK). The HDK is a USB bus-power board requiring a PC-host for power and interfacing. The HDK interfaces with the open-source Rockpool toolchain for deployment and testing. An analog microphone and a analog jack are provided for direct analog single-channel differential input. Encoded audio data can alternatively be streamed from the host PC. Inference is performed on the Xylo device (red outline).*

Hardware	Idle (mW)	Act. (mW)	Dyn. (mW)	Dyn. E (mJ/Inf)	Act. E (mJ/Inf)
GPU [1]	14970	37830	22860	29.67	49.1
CPU [1]	17010	28480	11470	6.32	15.7
Jetson [1]	2640	4980	2340	5.58	11.9
MOVIDIUS [1]	210	647	437	1.5	2.2
LOIHI [1]	29	110	81	0.27	0.37
LOIHI [11]	29	40	11	0.037	0.13
SpiNNaker2 [11]	—	—	7.1	0.0071	—
Xylo (ours)	0.216	0.507	0.291	0.0066	0.011

Table 2: *KWS task energy benchmarking in comparison with traditional and neuromorphic architectures. Power measured on physical devices in all cases. Act.: Active; Dyn.: Dynamic; E: Energy per inference. Active energy is not reported for SpiNNaker2 in the source benchmark paper [11], but this is elsewhere reported as* 390 mW *[6].*

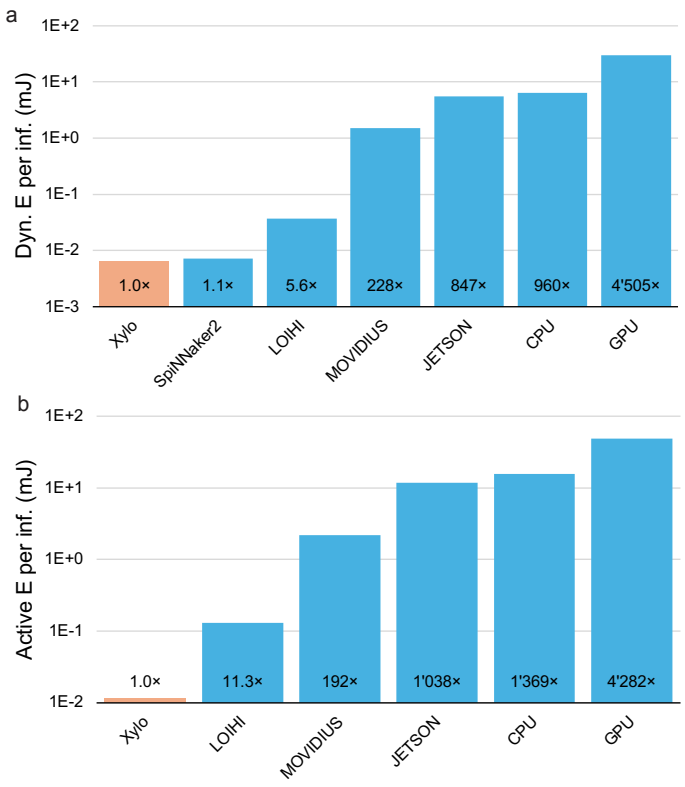

Figure 6: *Energy per inference comparison for the Aloha KWS benchmark task on several hardware architectures. a Dynamic energy per inference comparison. This is the standard metric reported for the Aloha benchmark. b Active energy per inference comparison. Active energy is not reported for SpiNNaker2 in the source benchmark paper [11], but this is elsewhere reported as* 390 mW *[6]. Energy per inference for Xylo defined as baseline* (1.0 ×)*. See Table 2 for precise values.*

Discussion

We implemented the Aloha spoken KWS benchmark task on Xylo Audio 2. Our trained network achieved high task accuracy despite its compact size, and the deployed quantised network suffered from only a small drop in accuracy ($<2\,\%$). We measured power used by the physical Xylo Audio 2 device while performing inference on the benchmark test set, and computed the inference rate for the system.

We found that Xylo Audio 2 exhibited high task accuracy (higher than the benchmark standard of 93%); performed inference faster than real-time ($>4\times$ speedup); and required 291 µW of dynamic power for inference. Xylo Audio 2 beat all other benchmarked devices on idle power, active power, dynamic power and inference efficiency.

The benchmark results reported here, as well as reported benchmarks for other hardware devices, do not include the power required for audio preprocessing. Most implementations of the Aloha benchmark require computation of an MFCC spectrogram, which can be computationally demanding. We used a simulation of the Xylo Audio 2 audio encoding block for audio preprocessing in simulation (Figure 2). We have measured the power consumed by the audio pre-processing block on Xylo Audio 2 as <50 µW.

Xylo Audio 2 is designed to operate as a real-time device for in- and near-sensor signal processing. Here we are operating the device in accelerated time, achieving a speed-up of $>4\times$, and an inference rate of >40 Hz. This is a lower inference rate than obtained for inference accelerator designs such as LOIHI, SpiNNaker2 and GPUs. These devices are designed to operate at high inference rates on large volumes of data, often making extensive use of parallel processing. In contrast with these systems, Xylo is designed to be an efficient real-time processor, operating on a continuous (i.e. non-batched) real-world signal. This is reflected by the energy efficient performance of Xylo at moderate inference rates.

Our results underscore the efficiency of Neuromorphic processor designs. Previous benchmark results for other Neuromorphic devices have shown large-factor gains in energy efficiency over low-power conventional processors [10], mobile inference processors and inference ASICs [3, 9, 8], commodity CPUs [1, 7, 9, 4] and GPUs [1, 7, 9, 4]. We show that Xylo Audio 2 sets a new standard for general-purpose Neuromorphic processors, exhibiting micro-power operation on continuous real-time signal processing tasks.

Bibliography

[1] Peter Blouw et al. *Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware*. en. arXiv:1812.01739 [cs, stat]. Apr. 2019. URL: `http://arxiv.org/abs/1812.01739` (visited on 04/20/2023).

[2] Hannah Bos and Dylan Muir. "Sub-mW Neuromorphic SNN Audio Processing Applications with Rockpool and Xylo". In: *Embedded Artificial Intelligence: Devices, Embedded Systems, and Industrial Applications*. CRC Press, 2022, pp. 69–78. ISBN: 978-87-7022-821-3. URL: `https://ieeexplore.ieee.org/book/9967439`.

[3] Charlotte Frenkel and Giacomo Indiveri. "ReckOn: A 28nm sub-mm2 task-agnostic spiking recurrent neural network processor enabling on-chip learning over second-long timescales". In: *2022 IEEE International Solid-State Circuits Conference (ISSCC)*. Vol. 65. IEEE, 2022, pp. 1–3.

[4] Dominique J. Kösters et al. "Benchmarking energy consumption and latency for neuromorphic computing in condensed matter and particle physics". In: *APL Machine Learning* 1.1 (Mar. 2023). Publisher: AIP Publishing, p. 016101. DOI: `10.1063/5.0116699`. URL: `https://doi.org/10.1063%2F5.0116699`.

[5] Dylan Muir, Felix Bauer, and Philipp Weidel. "Rockpool Documentaton". en. In: (2023). Publisher: Zenodo. DOI: 10.5281/ZENODO.3773845. URL: https://zenodo.org/record/3773845.

[6] Khaleelulla Khan Nazeer et al. *Language Modeling on a SpiNNaker 2 Neuromorphic Chip*. arXiv:2312.09084 [cs]. Jan. 2024. DOI: 10.48550/arXiv.2312.09084. URL: http://arxiv.org/abs/2312.09084 (visited on 06/17/2024).

[7] Christoph Ostrau et al. "Benchmarking Neuromorphic Hardware and Its Energy Expenditure." eng. In: *Frontiers in neuroscience* 16 (2022). Place: Switzerland, p. 873935. ISSN: 1662-4548 1662-453X. DOI: 10.3389/fnins.2022.873935.

[8] Fabrizio Ottati et al. *To Spike or Not To Spike: A Digital Hardware Perspective on Deep Learning Acceleration*. arXiv:2306.15749 [cs]. Jan. 2024. DOI: 10.48550/arXiv.2306.15749. URL: http://arxiv.org/abs/2306.15749 (visited on 06/15/2024).

[9] Gavin Parpart et al. *Implementing and Benchmarking the Locally Competitive Algorithm on the Loihi 2 Neuromorphic Processor*. arXiv:2307.13762 [cs]. July 2023. DOI: 10.48550/arXiv.2307.13762. URL: http://arxiv.org/abs/2307.13762 (visited on 06/14/2024).

[10] Georg Rutishauser et al. "7 μJ/inference end-to-end gesture recognition from dynamic vision sensor data using ternarized hybrid convolutional neural networks". In: *Future Generation Computer Systems* (2023). ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2023.07.017. URL: https://www.sciencedirect.com/science/article/pii/S0167739X23002704.

[11] Yexin Yan et al. "Comparing Loihi with a SpiNNaker 2 prototype on low-latency keyword spotting and adaptive robotic control". en. In: *Neuromorphic Computing and Engineering* 1.1 (July 2021). Publisher: IOP Publishing, p. 014002. ISSN: 2634-4386. DOI: 10.1088/2634-4386/abf150. URL: https://dx.doi.org/10.1088/2634-4386/abf150 (visited on 06/14/2024).

Key personnel

Dr. Hannah Bos, PhD — *ML Engineer*

Dr. Bos is a Senior Algorithms and Applications ML Engineer at SynSense, with a background in computational Neuroscience and theoretical Physics. At SynSense she designs algorithms for neuromorphic chips and helps with the design of new hardware. Dr. Bos holds a PhD in Physics and theoretical Neuroscience from RWTH Aachen, and a Masters in Physics from the University of Oslo.

Dr. Dylan Muir, PhD — *VP, Global Research Operations and Programme Manager*

Dr. Muir is a specialist in architectures for neural computation. He has published extensively in computational and experimental neuroscience. At SynSense he is responsible for directing development of neural architectures for signal processing. Dr. Muir holds a Doctor of Science (PhD) from ETH Zürich, and undergraduate degrees (Masters) in Electronic Engineering and in Computer Science from QUT, Australia.