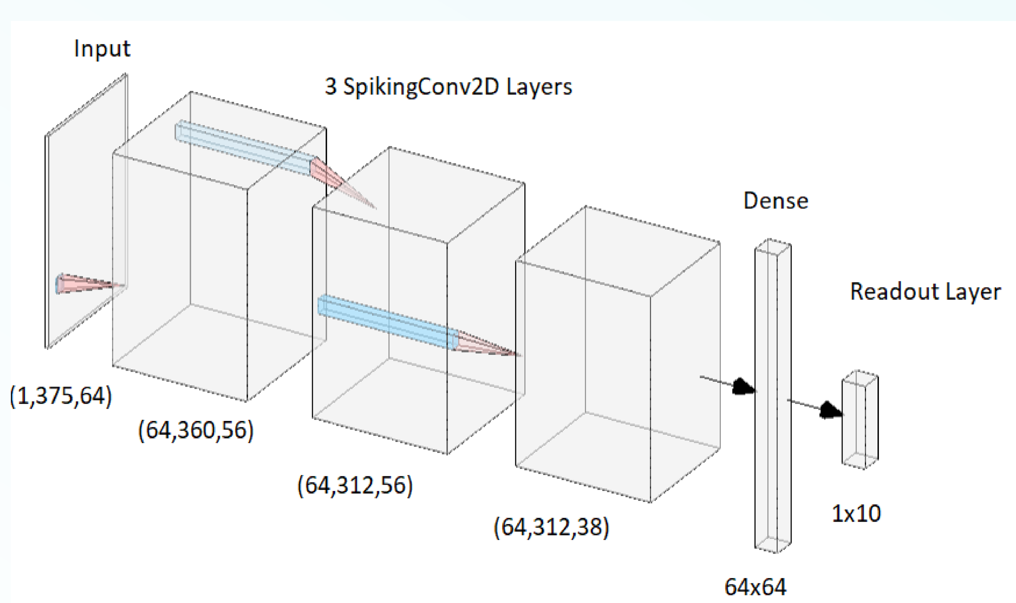
Classifying sounds using Dynamic Audio Sensors

Akar Adrien, Azzouz Firas, Bettaieb Ghassen,
Boussif Oussama, Hachicha Mohamed Amine

CentraleSupélec

Dynamic Audio Sensors

- Dynamic Audio Sensors (DAS) are silicon sensors inspired by the biological cochlea. [1]
- DAS are more energy efficient than standard microphones.
- The DAS has 64 audio frequency channels: each channel fires when there's a spike in the corresponding frequency.
- We want to evaluate Spiking Neural Networks ability to classify sounds recorded with DAS.



SNN Architecture (Zimmer et al 2019)

State of the art models summary

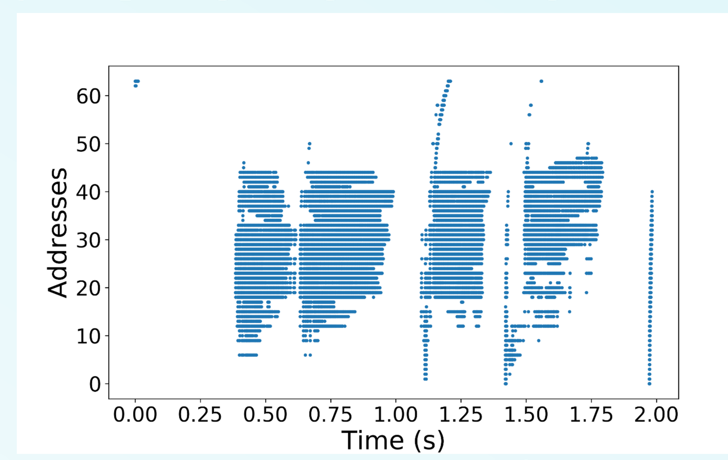
Feature type	Sensor	Task	Classifier	Accuracy (%)
MFCC		**Digit**	**GRU RNN**	**97.90**
Binned frames (fixed bins/sample)*	AMS1b	Digit	SVM	95.08
Constant time bins**	AMS1b	Digit	CNN	87.65
Constant time bins**	AMS1b	Digit	GRU RNN	82.82
Single events (raw data)	AMS1b	Digit	Phased LSTM	87.75
Data-driven time-binned features	AMS1b	Digit	Phased LSTM	91.25[a]
Constant time bins	AMS1b	Digit	GRU RNN	86.4
Exponential features	**AMS1b**	**Digit**	**GRU RNN**	**90.9**
Constant time bins	AMS1c	Digit	GRU RNN	88.6
Exponential features	**AMS1c**	**Digit**	**GRU RNN**	**91.1**
Constant time bins	AMS1b	Sequence	LSTM RNN	86.1[b]
Exponential features	**AMS1b**	**Sequence**	**LSTM RNN**	**87.3[b]**

Accuracy results for different SOTA models.

✓ **The highest achieved accuracy for binned representation of data (see preprocessing) was with the SVM classifier: 95,08% accuracy.**

N-TIDIGITS Spikes Dataset

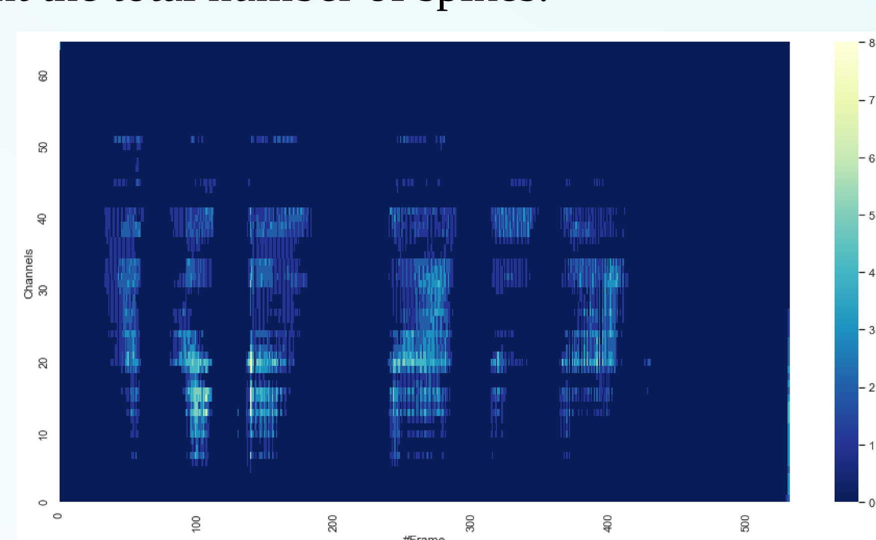
- 8621 recordings : 2463 single digit recordings and other are sequence of digits.
- We only keep single digit recordings: from 0 to 9.



Spectrogram a recording

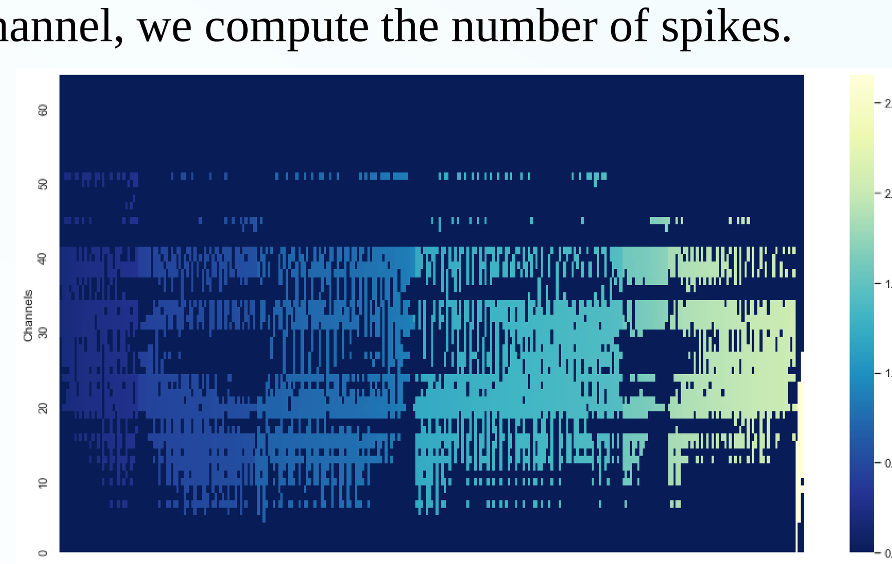
Preprocessing: Binning

Time binning : The recording is divided into equal time bins. For each time frame and for each frequency channel, we count the total number of spikes.



Time Binning of a recording

Event binning : The recording is divided into frames where for each frame we have the same total number of spikes E (all channels included). For each frame, and for each channel, we compute the number of spikes.



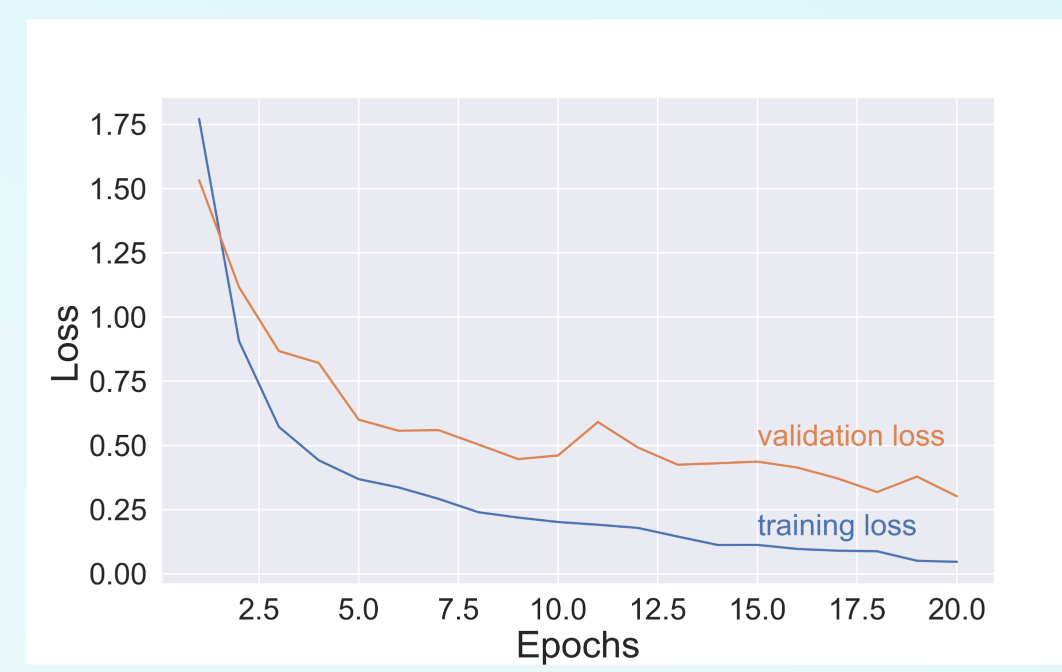
Event Binning of a recording

✓ **Event binning doesn't capture silence.**
✓ **Frames in event binning do not have the same duration.**

Training the SNN

Model parameters:

- Time binning with 5ms frame duration.
- 20 epochs (4 hours of training time).
- 3 hidden SpikingConv layers.
- We also tried with 2 SpikingConv Layers.

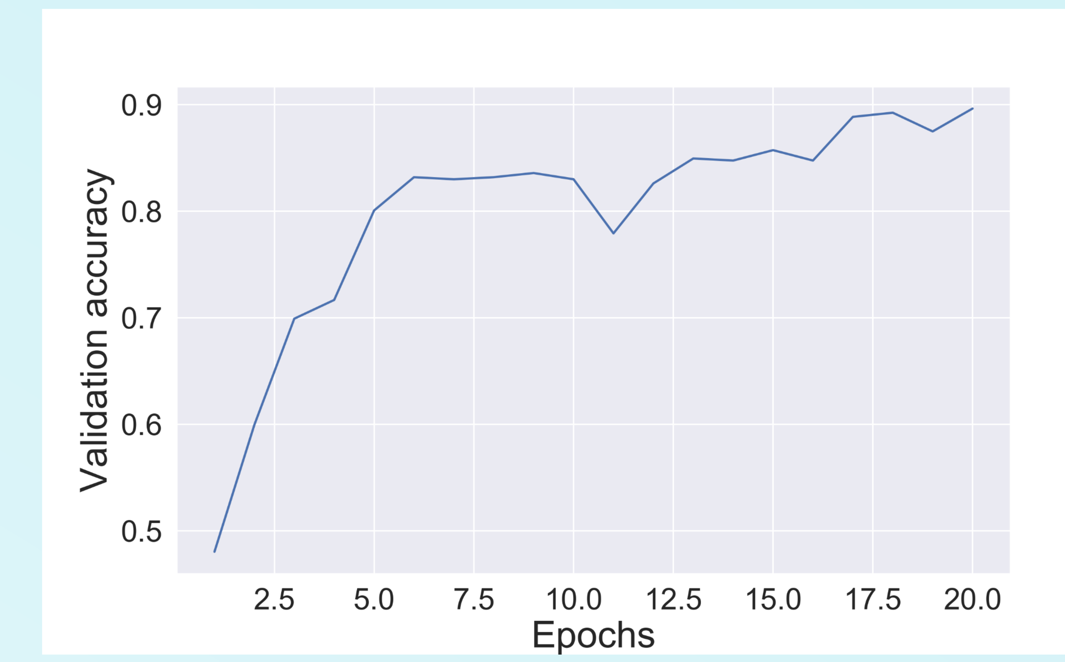


Train and validation loss evolution with time binning

The final SNN output for a batch of 9 samples

- **The final layer consists of 10 neurons: a neuron for each class.**
- **Each neuron outputs a signal and the dominant signal determines the class of the input signal.**

Classification Accuracy



Evolution of validation accuracy

- We reach a final validation accuracy of approximately 90% with time binning.
- Event binning fails to give satisfactory results.
- Simplifying the model (removing a SpikingConv layer for example) doesn't improve the performance.

Future directions

- Introduce dropout in the SNN to reduce the gap between training and validation loss.
- Normalize the Event and Time binning histograms.
- Segment the audio recordings made of digit sequences to increase the size of our training dataset of single digit recordings.

References

[1] Chan V, Liu SC, van Schaik A (2007) AER EAR: A matched silicon cochlea pair with address event representation interface.
[2] Anumula J, Neil D, Delbruck T, Liu SC (2018) Feature representations for neuromorphic audio spike streams.
[3] Zimmer R, Pellegrini T, Singh Fateh S, Masquelier T (2019) Technical report: supervised training of convolutional spiking neural networks with PyTorch.