EVENT-BASED LIP-READING WITH SPIKING NEURAL NETWORKS

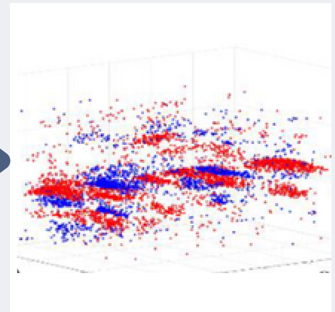
MADE BY: MARCEL SCHWEIKER (EIT DIGITAL), AND HUGO BULZOMI (M2 IA-ID)
SUPERVISED BY: JEAN MARTINET AND AMÉLIE GRUEL

EVENT CAMERA

Event camera
[Lichtsteiner,2008]

SPEAKER

EVENT VIDEO



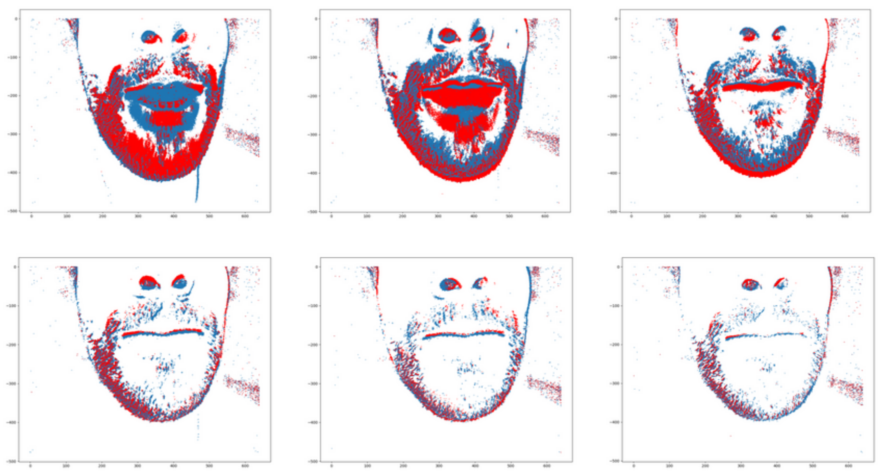
EVENT DATA ADVANTAGES

- **Resilient to motion blur**
- **High time resolution (microsecond)**
- **More compact with less redundant information (x, y, t, p)**
- **Power-efficient**
- **Works under low-light conditions**

PROJECT PIPELINE

PREPROCESSING TO VOXEL GRID

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



Results on **two datasets**:

- DVS-Lip *[Tan et al., 2022]*
- 2022_i3s_EventLipReadingDataset (i3s dataset) *[Pietrzak et Sabatier, 2022]*

Lip-Reading Model → WORD PREDICTION: "OYSTER"

COMPARISON OF TWO APPROACHES:

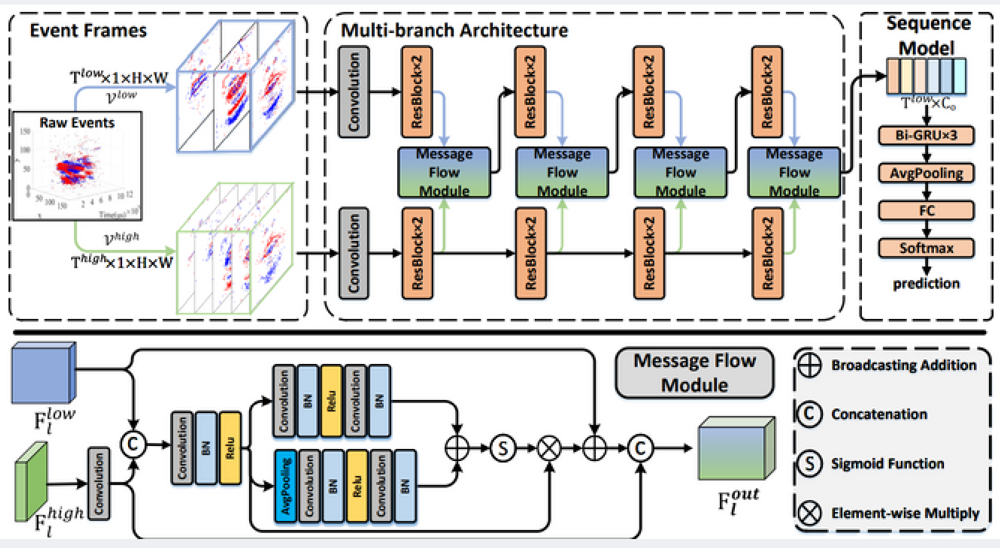
- **Regular deep ANN**
- **Using a deep SNN**

***[Lichtsteiner,2008]**: A128x128 120 dB 15 us Latency Asynchronous Temporal ContrastVision Sensor. IEEE Journal of Solid-State Circuits*
***[Tan et al, 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*
***[Pietrzak et Sabatier, 2022]**: Création d'un jeu de données de classification de données événementielles. Projet de TER Université Côte d'Azure*

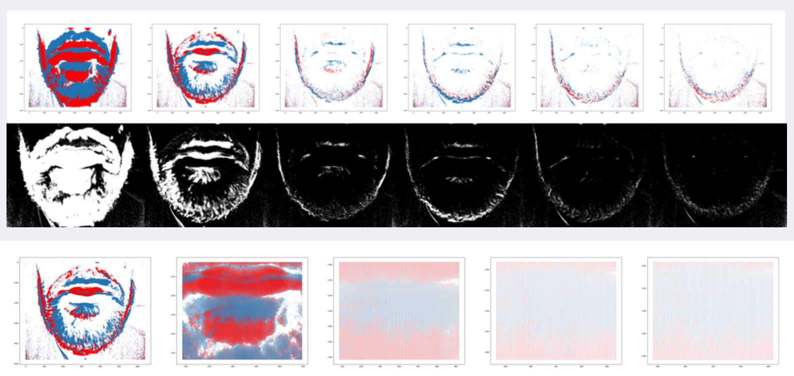
USING A DEEP ANN

Adaption of i3s dataset to MSTP [Tan et al., 2022]

- Rework Folder Structure and train-test split
- Transform numpy files to MSTP requirements
- Adjust centered crop to i3s data resolution
- Adapt hyperparametere, especially seq_len



Analysis of the i3s dataset



Shortcomings of the i3s dataset:

- non-centered mouth
- different distances to camera
- unnecessary spatial and temporal information
- resolution too high for MSTP

Proposed Experiments

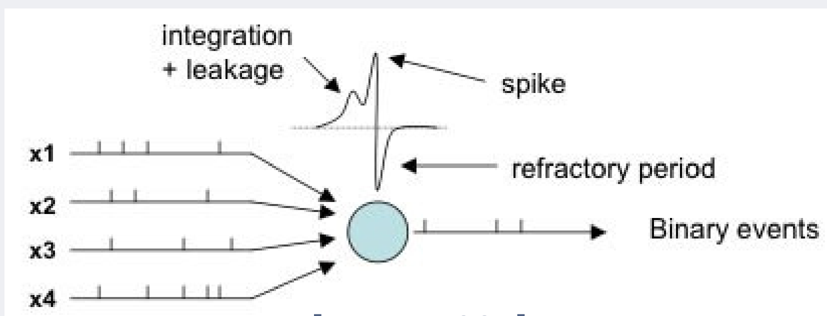
1. Augment input size of MSTP
2. Downscale i3s dataset using event count [Gruel et al., 2022]
3. Mouth Detection, mouth centered crop and uniform resizing of i3s dataset

Due to time and technical limitations, only downscaling experiments using different factors could be carried out.

***[Gruel et al., 2022]**: Event data downscaling for embedded computer vision. In 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*

USING A DEEP SNN

Spiking Neural Networks



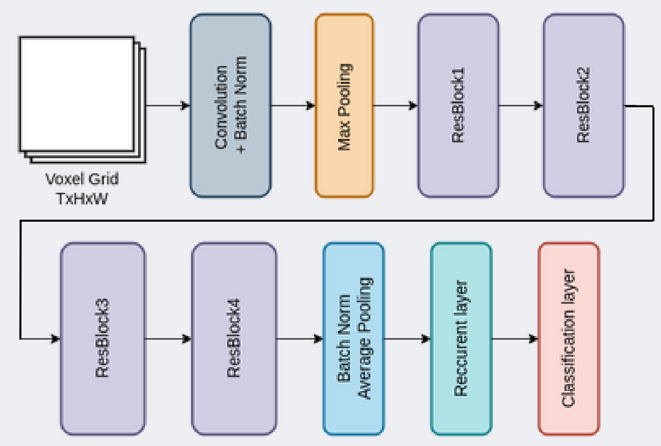
[Maas, 1997]

- Bio-inspired
- Energy efficient
- Theoricaly more expressive than regular neurons [Maas, 1997]
- Hard to train, as output from neuron is non-differenciable

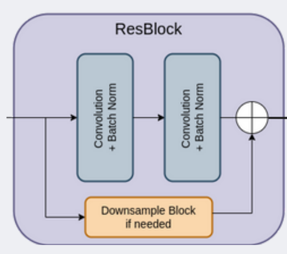
→ no gradient descent

We use a surrogate function to approximate the output of spiking neurons
→ surrogate gradient descent

Proposed SNN for Lip-Reading

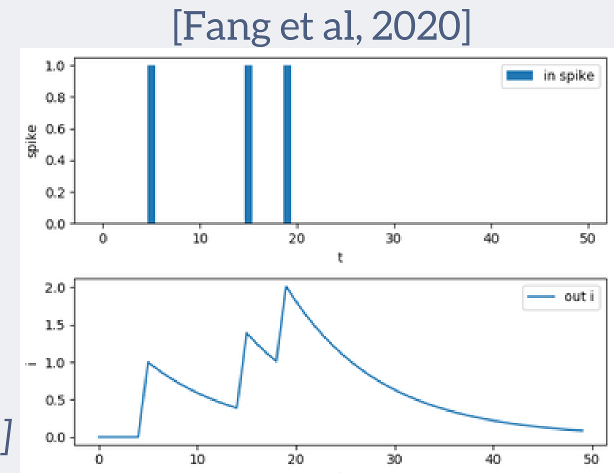


- Literature for video classification with SNN was lacking
- We propose the **first ever** deep SNN for lip-reading with a **spiking version** of the low-rate branch of **MSTP**



Original GRU layer of MSTP didn't work so well as a spiking layer. We replaced it by a **stateful synapse** that accumulate spikes and releases voltage slowly. *[Fang et al,2020]*

[Fang et al, 2020]



***[Fang et al, 2020]**: Exploiting neuron and synapse filter dynamics in spatial temporal learning of deep spiking neural network.*
***[Maas, 1997]**: Networks of Spiking Neurons: The third generation of neural network models. Neural Networks.*

Experiment Results

Experiments were carried out

- On a subset of 9 classes with a random train-test-split (see 'SDS ran' in Table 1)
- With a split based on individuals so that the model is evaluated on unseen participants (see 'SDS ind')
- On the whole dataset - 70 classes - with the individual split (see 'WDS ind')

Experiment	SDS ran	SDS ind	WDS ind
Initial State	0.852	0.617	0.330
Downscaling 2	0.938	0.567	0.378
Downscaling 3	0.975	0.555	0.411
Downscaling 4	0.914	0.444	-

Table 1: Overview of ANN Experiment results

RESULTS

Models	Dataset	Test Accuracy
MSTP	**DVS-Lip**	**0.721**
Spiking MSTP	DVS-Lip	0.602
MSTP	**i3s dataset**	**0.411**
Spiking MSTP	i3s dataset	0.081

Table 3: Results comparison between MSTP and spiking low-rate branch of MSTP on the DVS-Lip and i3s lip-reading datasets.

Experiment Results

We tried different methods for replacing the GRU layer, and show that the **stateful synapse** helps the network **remembering** past inputs.

The following table shows results gathered on the DVS-Lip dataset from [Tan et al., 2022].

Models	Experiment	Test Accuracy
MSTP	Tan et al. (2022)	**0.721**
Spiking MSTP	No GRU	0.522
Spiking MSTP	1 layer spiking GRU	0.463
Spiking MSTP	linear recurrent spiking neurons	0.476
Spiking MSTP	**stateful synapse**	**0.602**
Spiking MSTP	stateful synapses after layer	0.575

Table 2: Overview of SNN Experiment results