End-to-end Neuromorphic Lip Reading

Hugo Bulzomi, Marcel Schweiker, Amélie Gruel, Jean Martinet

End-to-end Neuromorphic Lip Reading

Hugo Bulzomi Marcel Schweiker Amélie Gruel Jean Martinet
i3S / CNRS, Université Côte d'Azur, Sophia Antipolis, France
Contact author: bulzomi@i3s.unice.fr

Abstract

Human speech perception is intrinsically a multi-modal task since speech production requires the speaker to move the lips, producing visual cues in addition to auditory information. Lip reading consists in visually interpreting the movements of the lips to understand speech, without the use of sound. It is an important task since it can either complement an audio-based speech recognition system or replace it when sound is not available. We introduce in this paper a neuromorphic model for lip reading, that uses events produced by an event-based sensor capturing lips motion as input, and that classifies short event sequences in word categories based on a SNN architecture. Experimental results show that the proposed model successfully leverages various advantages of neuromorphic approaches such as energy efficiency and low latency, which are central features in real-time embedded scenarios. To the best of our knowledge, it is the first proposal of an end-to-end neuromorphic lip reading model.

1. Introduction

Automatic lip reading is a computer vision task whose goal is to transcribe spoken words based solely on visual information. The prospect of systems capable to read lips is attractive for a variety of real-life applications, like assistive devices for persons with disabilities, security and surveillance applications, or automatic transcription of video content.

Traditional lip reading systems often rely on frame-based cameras to capture lip movements and require extensive processing to extract relevant information. Moreover, regular state-of-the-art methods for action recognition tend to be very computationally expensive and energy-demanding. Past years have thus seen growing interest surrounding cost-effective and portable approaches to solve visual tasks [11, 18]. In this context, research on neuromorphic technologies has skyrocketed, due to their intrinsic energy efficiency.

Event-based cameras, which capture motion information in a spatiotemporal fashion, have opened up new possibilities for developing energy-efficient and low-latency action recognition systems [32]. Along with this new generation of sensors, spiking neural networks (SNNs) have become very popular for their time-based processing, similar to how biological neurons communicate through spikes. The combination of event cameras and spiking neural networks has shown great potential in developing efficient and low-latency visual processing systems [4, 21], particularly when traditional cameras and computing methods struggle to meet the energy-efficiency requirements. In scenarios of battery-powered embedded devices, these technologies may prove to be incredibly useful. An end-to-end neuromorphic lip-reading system could directly benefit the development of portable assistive devices or surveillance equipment.

In this paper, we propose a neuromorphic model for lip reading that utilizes events produced by an event-based sensor capturing lip motion. The model classifies short sequences of events into word categories using a SNN architecture. The proposed model leverages the advantages of neuromorphic computing, including energy efficiency and low latency, making it suitable for real-time embedded scenarios. Our work draws inspiration from Tan et al. [33], who achieved state-of-the-art results with an original deep neural network architecture for an automatic lip reading task: the Multi-grained Spatio-Temporal Features Perceived (MSTP) network. Experimental results demonstrate the effectiveness of the proposed model in lip reading, achieving high accuracy in word classification. To the best of our knowledge, this is the first proposal of an end-to-end neuromorphic lip reading model. We believe it holds great potential for applications in human-computer interaction, speech recognition, and assistive technologies for the hearing-impaired.

2. State of the art

Automatic lip reading: Compared to other visual tasks, automatic-lip reading deals with especially subtle movements with very precise timing and is very sensitive to noise and other environmental factors. A model for automatic lip reading thus needs to excel at extracting both spatial and

temporal information. This makes lip reading a challenging task amongst other visual recognition problems (like face or object recognition), and lead to many different approaches throughout the years.

The first automatic lip reading system was made by [29] and used image video grayscale thresholding to try to match the contours of the face, nostrils, and mouth. This approach was very popular until the end of the 80s when [36] first applied a statistical classification model in the form of an early neural network to this problem. Image frames of size 20×25 were used and passed to a feed-forward neural network, and allowed to extract more information than the previous symbolic method. In later years, hidden Markov chains were applied by [19] and specifically used motions produced by the oral cavities region as features. Hidden Markov chains remained the most popular method throughout the 2000s, but as deep learning started gaining more traction in the mid-2010s, new works using artificial neural networks (ANNs) started being published.

In recent years, deep learning has been applied to automatic lip reading tasks, with some of the earliest examples dating back to the mid-2010s, e.g. [23] or [30]. The use of deep learning for lip reading has grown in popularity due to its ability to automatically learn and extract features from raw data, such as video frames of a person's face and mouth, without the need for manual feature engineering. This makes it well-suited for tasks like lip reading, where the exact movements of a person's mouth and lips are difficult to model and represent in a traditional, rule-based manner.

As discussed in Chung et al. [6], a big challenge for progress in the field of Deep Learning for lip reading has been the lack of suitable datasets. But with the availability of large amounts of data and the development of more powerful computing hardware, the use of deep learning has been enabled for these tasks. Existing lip reading datasets can be divided into several categories depending on the type of recognition object they are designed to capture. Alphabet recognition datasets like RMAV [9] or AV Letters [26] contain videos of individuals speaking the letters of the alphabet. Digit recognition datasets like OuluVS2 [3] and XM2VTSDB [28] are their equivalent for single digits. The LRW-1000 [34] dataset, as well as the AVSR [10] datasets, are examples of word recognition datasets. Finally, the last category is sentence recognition datasets, such as LRS3-TED [1], GRID [7], and LSVSR [31]. All of those datasets were recorded with RGB cameras. The quality and size of these datasets can vary greatly, with some containing thousands of videos and others containing only a few hundred.

Feng et al. [16] employ a convolutional neural network (CNNs) for visual feature extraction, then recurrent neural networks (RNN) to model long-term dependencies and better understand the context and meaning of the words being spoken. The combination of CNNs and RNNs is particularly useful for modeling the spatiotemporal nature of speech and accurately transcribing words that may span multiple frames of the video.

Event-based lip reading: Event cameras are a recent type of vision sensor that operates differently from traditional cameras [17]. Rather than capturing entire frames of images at fixed intervals, event cameras produce events, i.e., time-stamped pixel-level brightness changes, asynchronously and independently. Overall, the unique characteristics of the captured event-based data such as a high temporal solution, robustness to motion blur, asynchronous operation, and low power consumption are especially valuable for energy-efficient and embedded computer vision tasks like automatic lip reading. As discussed in [33], the existing event-based action recognition methods such as point-cloud-based, graph-based, and fixed-frame-based approaches are not suitable for the lip reading task, since it requires the perception of fine-grained spatiotemporal features from the event data.

Recently, Tan et al. [33] proposed a novel model architecture to perform event-based lip reading: the MSTP network. The performance of the MSTP on this task was experimentally proven to be superior to the state-of-the-art event-based action recognition models and video-based lip reading models. Within the MSTP network, the input events are consequently converted to low-rate and high-rate event frames with different temporal bins to preserve the spatiotemporal information of the event stream better. These two types of event frames are then fed into a multi-branch network with message flow modules between different branches designed to perceive both complete spatial features and fine temporal features from the event data. Followed by that, a sequence model decodes the visual features into words. In their study, SNNs are mentioned but discarded because of the lack of an efficient back-propagation algorithm to train them. Hence, SNNs are yet to be applied to this problem.

Spiking neural networks: SNNs are bio-inspired learning models that were first popularised in computer science by Wolfgang Maass [25]. Where regular ANNs are mathematical functions based on highly simplified brain dynamics, SNNs try to mimic the behaviour of biological neurons by emitting voltage "spikes" with precise timing. The most common spiking neuron model used is the Leaky-Integrate-and-Fire (LIF), which uses a parameter τ to adjust the speed at which the membrane potential will "leak" towards the resting potential. Parametric Leaky-Integrate-and-Fire neurons [15] (PLIF) make an interesting variation of regular LIF, where the time constant is adjusted during training. This means that the speed at which the membrane potential

of those neurons will leak is no longer a hand-tuned parameter. Their membrane potential evolves following Eq. 1:

$$V[t] = V[t-1] - \frac{1}{\tau}(V[t-1] - V_{reset}) + X[t] \quad (1)$$

with V[t] the membrane potential at time t, V_{reset} the resting potential, X[t] the neuron's input at time t, and τ the time constant that will be learned.

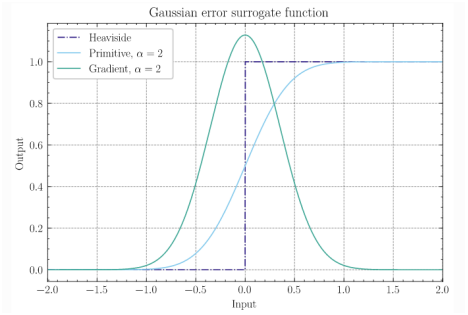
Gradient descent is impossible with SNNs because of the non-differentiable spike function of spiking neurons. The training of those networks is thus notoriously difficult. Until recent years, researchers have mainly been relying either on simple learning mechanisms (like STDP) that are often inefficient for supervised learning, or on the conversion of a regular pre-trained ANN into an SNN. More recent works however have been very successful in finding tricks and workarounds to be able to use gradient descent with SNNs. One of the most popular methods is described by Neftci et al. [27]: a differentiable surrogate function is used during training instead of the undifferentiable spike activation function. Amongst the most prominent ones, we can cite piecewise quadratic, ATan, and the Gaussian error surrogate function (Erf), shown in Fig. 1.

This surrogate function is an approximation of the spike function and allows the back-propagation of error through the network. Gradient descent is thus made possible with SNNs. This technique allowed deep spiking networks to equal, and sometimes surpass regular ANNs in recent works on real-life problems such as object recognition in automotive data [8]. Still, this method does not utilize the full potential of SNNs since it also brings some of the constraints that regular ANNs have (like the unidirectional flow of information through the network layers).

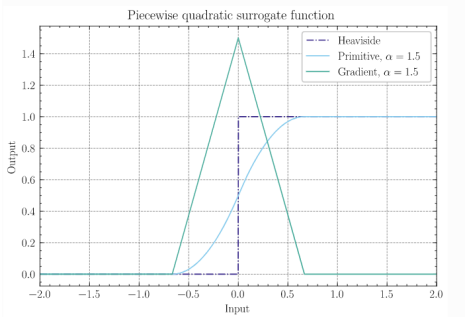
With the popularisation of surrogate gradient descent, efforts have been made to provide a spiking equivalent of temporal information extraction methods used in regular ANNs. Linearly recurrent spiking neurons are regular spiking neurons with a recurrent linear connection that makes their current output also depend on their previous one, in a manner similar to vanilla RNN. With a similar method, spiking-LSTM [24] has been proposed and declined into spiking Gated Recurrent Units (GRU). In another approach, stateful synapses [12] are placed after a spiking layer and provide additional memory by accumulating input spike current, making their output depend on both present spike input and previous ones. Fig. 2 shows an example of how those synapses behave given some spikes in the previous layer.

Those synapses can also be seen as leaky integrate neurons that instead of firing spikes, simply output their membrane potential.

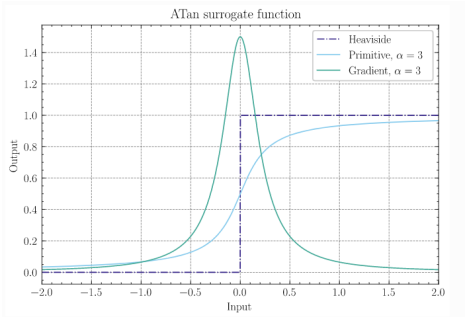
In visual applications, SNNs are a perfect match for the asynchronous chains of events produced by an event-based



(a) Erf



(b) Piecewise quadratic



(c) ATan

Figure 1. Some surrogate activation functions, from the SpikingJelly documentation [13].

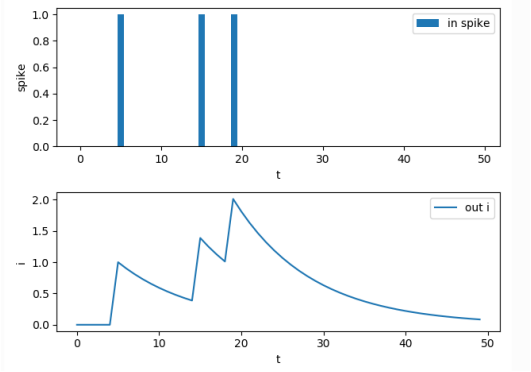


Figure 2. Output of the stateful synapse given some input spikes, from SpikingJelly documentation [13].

camera. The perspective of energy-efficient cameras and learning models to go with made SNNs very attractive for embedded systems like drones, or autonomous cars. SNNs have been successfully used for gesture recognition in numerous studies such as in Amir et al. [2], showing their potential for this type of task. As previously stated, very recent advances have been made in object recognition with event data like in the work of Cordone et al. [8]. Whereas SNNs were only marginally used on very simplistic tasks just a few years back, they are now able to equal the performances of regular ANNs on real-life problems.

As of now, SNNs are yet to be applied to automatic lip reading despite their attractiveness, supported by the arguments listed above. This innovative work is therefore the first to exploit its advantages for the specific task of lip-reading on event data.

3. Proposed methodology

As mentioned before, the asynchronous nature of SNN should make them a perfect match for the event data produced by event cameras. In the context of surrogate gradient descent though, it becomes necessary to convert our data to a synchronous form.

3.1. Event data preprocessing

Finding efficient ways to represent event data is a difficult problem that is, in itself, a subject of prior studies (e.g. [20]). Since we target the DVS-Lip dataset, we chose to use a method similar to the one described by [33]. In their study, the authors converted the asynchronous events into a 3-dimensional array, i.e. a voxel grid. Each event in our dataset is represented as an (x, y) position on the sensor, a time (t), and a polarity {-1, 1}. Equations 2 and 3 (reused from [33]) show how we can create a voxel grid of a specific length T where the polarity of each event is spread through the two closest spatiotemporal voxels.

$$t_k^* = \frac{T-1}{t_N - t_1}(t_k - t_1) \tag{2}$$

$$V(t,y,x) = \sum_k p_k max(0, 1 - |t - t_k^*|) \tag{3}$$

where T is the number of frames we want to use (i.e. the time resolution of our grid), and t_x is the timestamp of the x^{th} event from the original video. With our event data discretized this way, we get a 3-dimensional grid of shapes (t, x, y). In [33], Tan et al. use two different values for T: 30 for the low-rate branch of their network, and 210 for the high-rate branch. However, the individual performances of each branch when not combined are very similar (accuracy of 69.57% and 69.49% respectively for the low and high branches). Only when both are combined in a multi-grained network along with message flow module blocks do the overall accuracy increases to 72.10%. We thus decided to experiment using $T = 30$, since using a higher value would tremendously slow the training processes for little improvement.

3.2. Topology exploration

To the best of our knowledge, no prior work uses SNNs to classify dynamic scenes for lip reading. Moreover, the literature on dynamic classification with SNNs is scarce. The closest studies to our work used the DVS-Gesture dataset introduced by [2] along with reasonably simple topologies and different variations, like in Yao et al. [35]. The DVS-Gesture dataset itself being comparatively easy to classify, we hope to contribute to the area of neuromorphic computer vision by showing that more difficult problems can be tackled with SNNs.

We tested several topologies of different levels of complexity. A good starting point is to simply reuse some simple topologies proposed in prior work trying to classify DVS-Gesture, or for other visual tasks such as event-video reconstruction [37]. We have used the topology of Yao et al. [35]. Fig. 3 shows two simple SNN topologies that we used; we will refer to those models as SNN1 and SNN2.

SNN1 is borrowed from [35], who originally designed this topology to classify DVS-Gesture. SNN2 is inspired by Zhu et al. [37], where the authors use a spiking encoder-decoder architecture for event-video reconstruction. SNN2 is the encoder part of their model, where the residual layers have been replaced by two more convolution layers, yielding a higher accuracy. Batch normalization layers have been added after each convolution since such layers have been shown to considerably facilitate the learning process. On this topic, [8] showed batch normalization layers to be crucial when using complex SNNs, and reported either significant performance drop, or networks simply not learning when not using batch normalization.

Along with these simple models, we designed a spiking equivalent of the low-rate branch of MSTP. Since this model uses ResNet [22] as the backbone, we first implemented a spiking ResNet backbone, in a similar way as Fang et al. [14] did, and then applied the MSTP architecture. The topology of our spiking MSTP low-rate branch is presented in Fig. 4. Compared to the previous basic SNNs presented, there was no need to add batch normalization layers here, since the base model already feature them. Overall, little modifications have been made to the original model, and the resulting spiking low-rate branch MSTP ended up being very similar to the Spiking Element-Wise ResNet (SEW-ResNet) presented in [14].

One key difference between our spiking adaptation of MSTP and the original model lies in the GRU layer em-

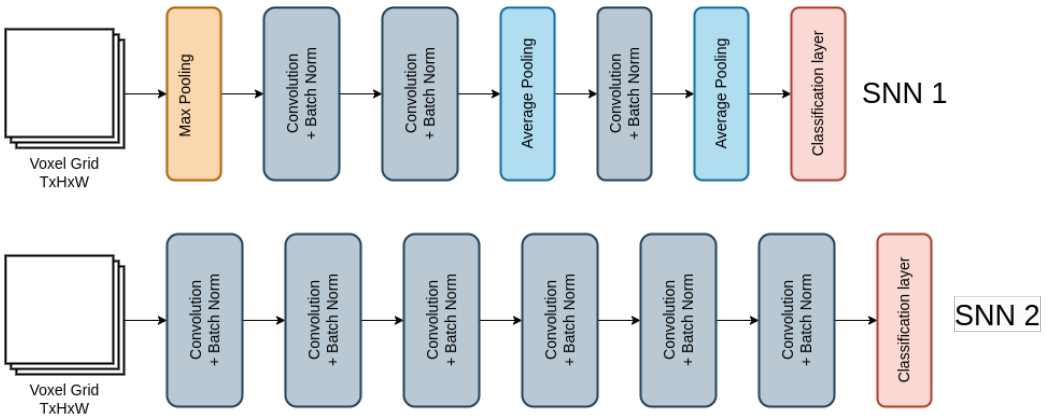


Figure 3. Two simple SNN topologies used in this work.

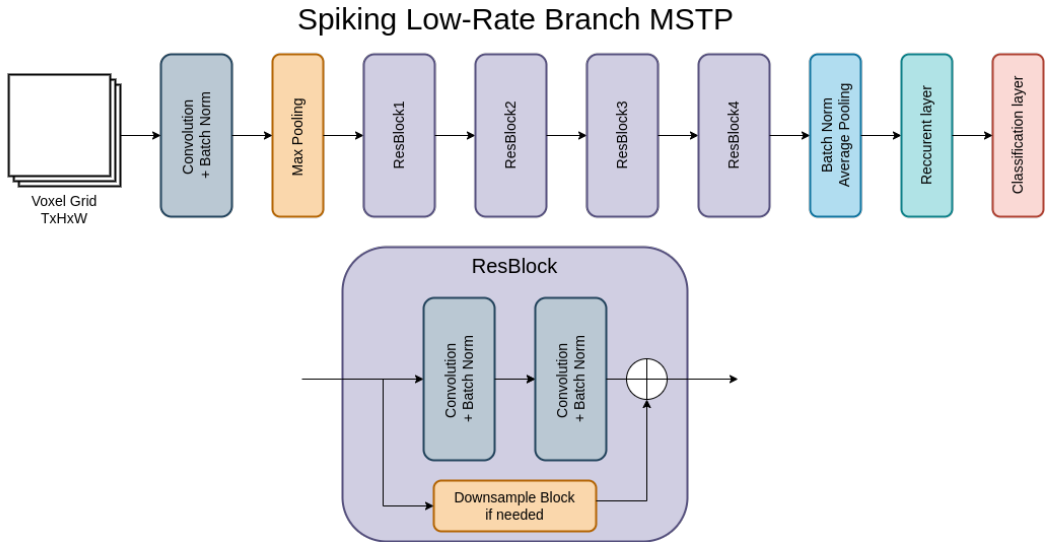


Figure 4. Overall spiking low-rate branch MSTP topology along with the ResBlock.

ployed near the output the architecture. RNNs in general, are typically employed when information has to be extracted from temporal sequences. Their output depends on both their current input as well as their hidden state. GRU in particular has been introduced by Chung et al. [5] and allows each neuron to learn how much of the previous information needs to be forgotten, and how much of the current new input should be memorized. Though a spiking LSTM layer (adaptable into a spiking GRU) has been proposed by Rezaabad et al. [24], we found it to give very underwhelming results in our case. Especially when trying to stick to the original 3 layers of bidirectional GRU used in MSTP, our network exhibited a very slow and inefficient learning process, and we ultimately decided to try to find alternatives for extracting temporal information. In the next section, we will describe in further details what techniques we tried using to replace this GRU layer.

3.3. Experiments

Experiments with SNNs were performed on the DVS-Lip dataset using 30 timesteps ($T = 30$ in Eq. 2) and were meant to help us find the most promising spiking topology before comparing MSTP to our best SNN. We used Parametric Leaky-Integrate-and-Fire neurons [15] (PLIF) with Adam optimizer with a learning rate of $1e^{-3}$, and a cosine annealing scheduler to adjust the learning rate during training in a manner similar to the work of Cordone et al. [8]. Neurons parameters are presented in Tab. 1 Using these settings, we ran experiments focused on the following 3 points.

Type	Initial τ	Activation threshold	Reset potential
PLIF	2.0	1.0v	0.0v

Table 1. Neurons parameters.

1 – First, we compared the previously introduced surrogate activation function in a small-scale experiment. As introduced earlier, surrogate gradient descent requires choosing an activation function to replace the regular non-differentiable step function used by spiking neurons. To select the most promising one, we ran fast test runs on a sub-section of 10 words of the DVS-Lip dataset using our smallest network (SNN1) for 50 epochs.

2 – Then, we compared the performances of each of the presented topologies, using the surrogate function that lead to the best results during the previous experiment. Even though we suspected our spiking MSTP branch to perform the best, we still wanted to try basic models to establish a spiking baseline. We thus trained each of the presented topologies for 100 epochs on the full DVS-Lip dataset.

3 – Finally, we conducted an ablation study to adapt the GRU layer from MSTP in our spiking adaptation. As mentioned, using a 3 layers bidirectional spiking GRU adapted from [24] yields underwhelming results. The point of using GRU is to help extract temporal information, but other methods to do this exist. We tested those we previously introduced: linearly recurrent spiking neurons and stateful synapses. We also tried to replace the GRU layer with a simple fully connected spiking layer. Each of those replacements was also tested for 100 epochs.

All experiments are run on a laptop with an Intel Core i9-12950HX CPU (2.5 GHz x 16), 62,5 GB RAM, with an NVIDIA RTX A5500 laptop GPU with 16 GB of VRAM.

4. Experimental results

4.1. Topologies Results

We discuss here the experimental results using various SNNs topologies and hyperparameters for classifying the DVS-Lip dataset. First, Tab. 2 shows the results of our first batch of training, which were meant to help us choose a surrogate activation function to use for the rest of the work.

Models	Activation function	Accuracy
SNN1	Erf	0.546
SNN1	Piecewise	0.531
SNN1	ATan	0.534

Table 2. SNN1 accuracy on a subset of 10 classes from DVS-Lip, using different surrogate functions. The classes correspond to the words: allow, allowed, America, American, benefit, benefits, billion, called, challenge, and change.

Models	Variation	Accuracy
SNN1	Base model	0.395
SNN2	Base model	0.514
Spiking MSTP	No GRU	0.522

Table 3. SNN results on the entire DVS-Lip dataset.

The choice of surrogate functions can vastly influence how our networks perform. This preliminary test shows Erf to perform slightly better than ATan and Piecewise, and we thus kept using it for the rest of our experiments.
After these preliminary tests, we then tested the 3 topologies presented earlier to see which one performs the best. Tab. 3 shows their respective performances.

This second batch of experiments shows that the spiking MSTP obtains significantly better results than the other models (SNN1 and SNN2).

4.2. Ablation study results

At this stage, the spiking MSTP does not use any recurrent layer, and the original GRU was simply replaced by a spiking Fully Connected layer. We have experimented with possible GRU replacement in order to see if a higher performance could be gained by using other temporal information extraction methods, as described in the previous section. The results of our third batch of experiments, and the final accuracy for the DVS-Lip Dataset using SNNs are presented in Tab. 4.

This table shows that significant performance growth can be gained by using stateful synapses either as a spiking replacement for the GRU layer. We also included the accuracy of the original MSTP published in [33], which is still higher than our best spiking model. Tab. 5 however, shows the accuracy and size of MSTP, the low-rate branch of MSTP only, and the low-rate branch of MSTP without its three-layer bidirectional GRU.

In this last table, we can observe that even though our model needs less than 5 times the amount of memory of

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**

Table 4. Spiking MSTP results on the entire DVS-Lip dataset, trying different substitutions for the 3-layer bidirectional GRU in the classification part.

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

Table 5. Size and accuracy comparison between our models and the state-of-the-art.

MSTP, it still manages to keep 83% of its accuracy. Furthermore, the main topological difference between our model and the low-rate branch of MSTP is the absence of the 3 layers of bidirectional GRU. If removed, the low-rate branch of MSTP is now the same size as our model but with slightly lower accuracy. In the end, the spiking nature of our model could make it an interesting option for embedded systems where our lower accuracy could still be seen as a good trade-off for a more energy-efficient model.

5. Conclusion

This paper proposes the first SNN model for event-based lip reading and presents competitive results with the current ANN state-of-the-art. We showed how we tested several topologies with various surrogate functions and improved our base results with a stateful synapse to extract more temporal information. Our model is amongst the first advanced deep spiking models applied to such a challenging task and manages to get promising results while keeping a relatively small size.

Further experiments to improve the data pre-processing and the SNN model itself may, in the near future, allow a state-of-the-art model to be achieved. In the absence of published deep and complex SNNs for similar tasks, we hope to provide a spiking baseline for future work in this area.

We think that improvements can still be made on our final SNN, especially regarding the way we replaced the GRU layer from MSTP. Our stateful synapses allowed us to break the 60% accuracy line, but we still think that most of the performance gap between MSTP and this model lies in this replacement. Finding more efficient ways to extract information from the temporal component of the data may be a good direction for future work. Furthermore, even though we believe surrogate gradient descent is the current best training method for supervised learning with SNNs, the results can be improved. By forcing gradient descent this way, we bring a lot of the limitations of regular ANN to our SNNs, while also using a lot of approximations during training. We thus hope that other training methods will be developed in the future, allowing us to tap into the full potential of SNNs.

Although the accuracy of the proposed model is lower than the current state-of-the-art, we still show that SNNs can be used for complex video classification tasks, since lip reading remains very challenging even for humans. Moreover, our work provides valuable insights for future studies in this area, as we proposed the first SNN model for automatic lip reading. Our work also shows that surrogate gradient descent does provide a worthwhile option for supervised training of SNNs, and our final spiking model shows the potential to have competitive results with those of a regular deep ANN.

Acknowledgements

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

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496, 2018. 2

[2] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017. 4

[3] Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1:1–5, 2015. 2

[4] Xuena Chen, Li Su, Jinxiu Zhao, Keni Qiu, Na Jiang, and Guang Zhai. Sign language gesture recognition and classification based on event camera with spiking neural networks. *Electronics*, 12(4):786, 2023. 1

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 5

[6] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. *CoRR*, abs/1611.05358, 2016. 2

[7] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 2

[8] Loïc Cordone, Benoît Miramond, and Philippe Thierion. Object detection with spiking neural networks on automotive event data. *arXiv preprint arXiv:2205.04339*, 2022. 3, 4, 5

[9] Stephen J Cox, Richard W Harvey, Yuxuan Lan, Jacob L Newman, and Barry-John Theobald. The challenge of multispeaker lip-reading. In *AVSP*, pages 179–184. Citeseer, 2008. 2

[10] Andrzej Czyzewski, Bozena Kostek, Piotr Bratoszewski, Jozef Kotus, and Marcin Szykulski. An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, 49, 10 2017. 2

[11] Yunbin Deng. Deep learning on mobile devices: a review. In *Mobile Multimedia/Image Processing, Security, and Applications 2019*, volume 10993, pages 52–66. SPIE, 2019. 1

[12] Haowen Fang, Amar Shrestha, Ziyi Zhao, and Qinru Qiu. Exploiting neuron and synapse filter dynamics in spatial temporal learning of deep spiking neural network. *arXiv preprint arXiv:2003.02944*, 2020. 3

[13] Wei Fang, Yanqi Chen, Jianhao Ding, Ding Chen, Zhaofei Yu, Huihui Zhou, Timothée Masquelier, Yonghong Tian, and

other contributors. Spikingjelly. https://github.com/fangwei123456/spikingjelly, 2020. Accessed: YYYY-MM-DD. 3

[14] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34:21056–21069, 2021. 4

[15] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2661–2671, 2021. 2, 5

[16] Dalu Feng, Shuang Yang, Shiguang Shan, and Xilin Chen. Learn an effective lip reading model without pains. *CoRR*, abs/2011.07557, 2020. 2

[17] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. 2

[18] Abhinav Goel, Caleb Tung, Yung-Hsiang Lu, and George K Thiruvathukal. A survey of methods for low-power deep learning and computer vision. In *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, pages 1–6. IEEE, 2020. 1

[19] Alan J Goldschen, Oscar N Garcia, and Eric D Petajan. Continuous automatic speech recognition by lipreading. In *Motion-Based recognition*, pages 321–343. Springer, 1997. 2

[20] Amélie Gruel, Jean Martinet, Teresa Serrano-Gotarredona, and Bernabé Linares-Barranco. Event data downscaling for embedded computer vision. In *17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, Online, Portugal, Feb. 2022. 4

[21] Jesse Hagenaars, Federico Paredes-Vallés, and Guido De Croon. Self-supervised learning of event-based optical flow with spiking neural networks. *Advances in Neural Information Processing Systems*, 34:7167–7179, 2021. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[23] Oscar Koller, Hermann Ney, and Richard Bowden. Deep learning of mouth shapes for sign language. volume 2015-, pages 477 – 483. IEEE, 2015. 2

[24] Ali Lotfi Rezaabad and Sriram Vishwanath. Long short-term memory spiking networks and their applications. In *International Conference on Neuromorphic Systems 2020*, pages 1–9, 2020. 3, 5, 6

[25] Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. 2

[26] I Matthews, T Cootes, J Bangham, S Cox, and R Harvey. Extraction of visual features for lipreading. *IEEE Trans. on Pattern Analysis and Machine Vision*, 24(2):198–213, 2002. 2

[27] Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019. 3

[28] Eric K. Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N. Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:II–2017–II–2020, 2002. 2

[29] Eric Petajan. Automatic lipreading to enhance speech recognition. *Proc. CVPR'85*, 1985. 2

[30] Stavros Petridis and Maja Pantic. Deep complementary bottleneck features for visual speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 2304–2308. IEEE Press, 2016. 2

[31] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior, and Nando de Freitas. Large-scale visual speech recognition. 2018. 2

[32] Deepak Singla, Soham Chatterjee, Lavanya Ramapantulu, Andres Ussa, Bharath Ramesh, and Arindam Basu. Hynna: Improved performance for neuromorphic vision sensor based surveillance using hybrid neural network architecture. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2020. 1

[33] Ganchao Tan, Yang Wang, Han Han, Yang Cao, Feng Wu, and Zheng-Jun Zha. 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*, pages 20094–20103, 2022. 1, 2, 4, 6

[34] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. *CoRR*, abs/1810.06990, 2018. 2

[35] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10221–10230, 2021. 4

[36] Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71, 1989. 2

[37] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3594–3604, 2022. 4