# SPEECH RECOGNITION USING BIOLOGICALLY-INSPIRED NEURAL NETWORKS

*Thomas Bohnstingl*$^\star$, *Ayush Garg*$^\star$, *Stanisław Woźniak*$^\star$,
*George Saon*$^\dagger$, *Evangelos Eleftheriou*$^\star$ and *Angeliki Pantazi*$^\star$

$^\star$ IBM Research, Zurich
$^\dagger$ IBM Research AI, Yorktown Heights, USA

## ABSTRACT

Automatic speech recognition systems (ASR), such as the recurrent neural network transducer (RNN-T), have reached close to human-like performance and are deployed in commercial applications. However, their core operations depart from the powerful biological counterpart, the human brain. On the other hand, the current developments in biologically-inspired ASR models lag behind in terms of accuracy and focus primarily on small-scale applications. In this work, we revisit the incorporation of biologically-plausible models into deep learning and enhance their capabilities, by taking inspiration from the brain's diverse neural and synaptic dynamics. In particular, we propose novel deep learning units by introducing neural connectivity concepts emulating the axo-somatic and the axo-axonic synapses and integrate them into the RNN-T architecture. We demonstrate for the first time that such a model can yield performance levels competitive to the state-of-the-art. Moreover, our implementation has a significantly reduced computational cost and a lower latency.

***Index Terms—*** speech recognition, RNN-T, spiking neural networks, synapse types, spiking neural unit

## 1. INTRODUCTION

Recently, researchers working on machine learning (ML) have increasingly resorted to deep learning approaches [1, 2, 3, 4] for automatic speech recognition (ASR), where three special kinds of architectures have been widely adopted [5, 6]: architectures based solely on recurrent neural networks (RNNs), such as the RNN transducer (RNN-T) employing long short-term memory (LSTM) units [7], architectures based on RNNs with the attention mechanism, such as the listen attend and spell model (LAS) [8], and recently also transformer-based models, such as the conformer [9]. While transformer-based and attention-based models potentially provide higher accuracy, the RNN-T architecture exhibits a lower latency, allows for real-time transcription, and is even commercially deployed, for example in mobile devices [10]. Recently, researchers have also tried to merge the RNN-T and LAS models in order to combine the benefits of both [11, 12, 13].

Despite their great successes, all the aforementioned models take inspiration from biology only remotely and depart from the way speech recognition is performed in the brain. Researchers working on biologically plausible spiking neural networks (SNNs), have tackled ASR using a different strategy and mainly employed the leaky integrate-and-fire (LIF) neurons [14, 15, 16]. However, there are several limitations of these works. For example, they often use simpler network architectures compared to the state-of-the-art. Moreover, often only parts of those architectures employ biologically-inspired units and the learning algorithms used are inferior to error-backpropagation [17, 18]. Thus, the performance of those approaches lags behind their ML-based counterparts in terms of accuracy.

In this paper, we address these limitations by proposing an RNN-T architecture that incorporates biologically-inspired neural units. Specifically, we leverage the diverse neuron and synapse types observed in the brain and propose novel extensions of the LIF model with much richer dynamics. We build upon the spiking neural units (SNUs) [19], which allows us to leverage the advanced optimization capabilities from the ML domain [20, 21], that are essential for training speech recognition models [22]. We demonstrate that a state-of-the-art network architecture incorporating biologically-inspired units can yield competitive performance levels in a large-scale speech recognition application, while significantly reducing the computational cost as well as the latency. In particular, our key contributions are:

- a novel neural connectivity model emulating the axo-somatic synapses that enhances the threshold adaptation in biologically-inspired neurons,
- a novel neural connectivity model emulating the axo-axonic synapses that modulates the output of biologically-inspired neurons,
- a biologically-inspired RNN-T architecture, where the LSTM units have been replaced with SNU variants, resulting in a significantly reduced computational cost and an increased throughput, which still provides competitive performance compared to its LSTM-based counterpart.

## 2. BIOLOGICALLY-INSPIRED MODELS

Typically, architectures composed of biologically-inspired neurons consider only axo-dendritic synapses, where the axon of the presynaptic neuron is connected to the dendrite of the postsynaptic neuron, and the output from the presynaptic neuron is modulated by the synaptic weight. However, neural networks in the brain exhibit a much more complex connectivity structure [23, 24]. In this work, we leverage this diversity and investigate two additional types of synapses, namely the axo-somatic and the axo-axonic synapses, and propose novel biologically-inspired models with richer dynamics. Figure 1a shows an illustration of such neuron models and their connectivity via the various types of synapses.

The simplest biologically-inspired model is based on the LIF dynamics that can be abstracted into the form of a recurrent unit using the SNU framework [26]. Figure 1b shows the basic configuration

Correspondence to boh@zurich.ibm.com

E. Eleftheriou is currently with Axelera AI – Zurich

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

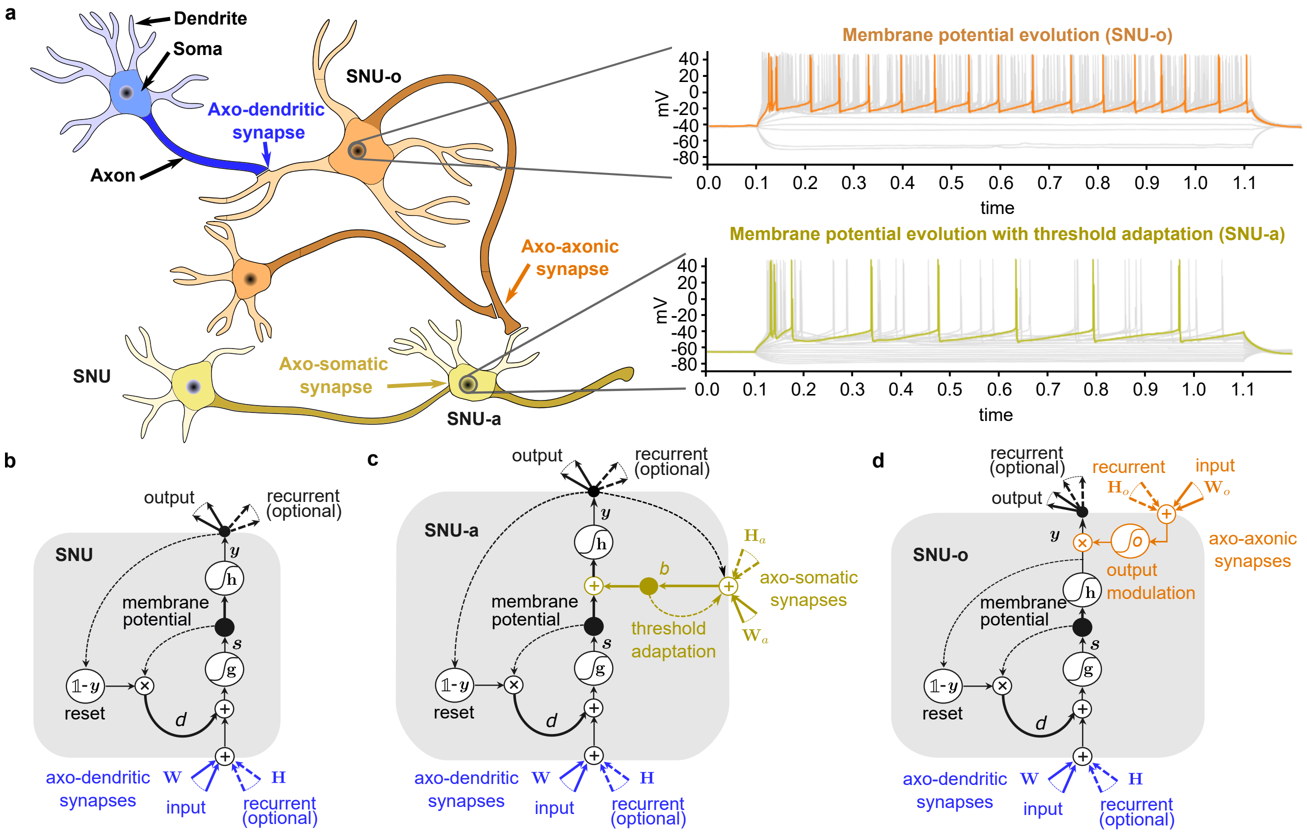

**Fig. 1**. **Illustration of different biological synapse and neuron types along with their realization in simulations**. **a** A neural network example composed of three neuron and three synapse types along with recordings of membrane potentials from the Human Brain Atlas [25]. **b** Visualization of SNU modelling the LIF dynamics using solely axo-dendritic synapses. **c** Visualization of the SNU-a modelling the adaptive threshold behavior using axo-somatic synapses, illustrated with yellow color in **a**. **d** Visualization of the SNU-o using axo-axonic synapses, illustrated in orange in **a**.

of the SNU. A layer of $n$ SNUs receiving $m$ inputs is governed by the following equations:

$$\boldsymbol{s}^t = \mathbf{g}(\mathbf{W}\boldsymbol{x}^t + \mathbf{H}\boldsymbol{y}^{t-1} + d \cdot \boldsymbol{s}^{t-1} \odot (\mathbb{1} - \boldsymbol{y}^{t-1})), \tag{1}$$

$$\boldsymbol{y}^t = \mathbf{h}(\boldsymbol{s}^t + \mathbf{b}), \tag{2}$$

where $\boldsymbol{x}^t \in \mathbb{R}^m$ represents the inputs at time step $t$, $\boldsymbol{s}^t \in \mathbb{R}^n$ represents the membrane potentials of the neurons, $\boldsymbol{y}^t \in \mathbb{R}^n$ represents the outputs of the neurons, $d \in \mathbb{R}$ represents the constant decay of the membrane potential, $\mathbf{b} \in \mathbb{R}^n$ represents the trainable firing threshold, $\mathbf{W} \in \mathbb{R}^{n \times m}$ and $\mathbf{H} \in \mathbb{R}^{n \times n}$ denote the trainable input and the recurrent weight matrices. Note that the SNU layer employs only axo-dendritic synapses, i.e., $\mathbf{W}$ and $\mathbf{H}$, indicated by the dark blue color in Fig. 1a. As described in [19], the SNU can operate in principle in two different modes, one in which the SNU yields continuous outputs (called sSNU), i.e. $\mathbf{h} = \boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ denotes the sigmoid function, and one in which the SNU emits discrete-valued spikes, i.e. $\mathbf{h} = \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ indicates the Heaviside step function. In this work, we focus on the dynamics of the neuronal models, in particular on the different synapse types, and thus mainly consider the sSNU variants.

One well-known property of neurons in the human brain is that they can adapt their firing threshold across a wide variety of timescales [25, 26, 27]. In biology, there are complex mechanisms influencing the firing threshold. In particular, it can be increased or decreased following the neuron's own dynamics, but also based on the activity of the other neurons [28, 29]. We propose a novel adaptive SNU (SNU-a) that enhances the threshold adaptivity dynamics by emulating the axo-somatic synapses, indicated with yellow color in Fig. 1a. A layer of $n$ SNU-a units is governed by

$$\boldsymbol{s}^t = \mathbf{g}(\mathbf{W}\boldsymbol{x}^t + \mathbf{H}\boldsymbol{y}^{t-1} + d \cdot \boldsymbol{s}^{t-1} \odot (\mathbb{1} - \boldsymbol{y}^{t-1})), \tag{3}$$

$$\boldsymbol{b}^t = \boldsymbol{\rho} \odot \boldsymbol{b}^{t-1} + (1 - \boldsymbol{\rho}) \odot \left(\mathbf{W}_a\boldsymbol{x}^t + \mathbf{H}_a\boldsymbol{y}^{t-1}\right), \tag{4}$$

$$\boldsymbol{y}^t = \mathbf{h}(\boldsymbol{s}^t + \beta\boldsymbol{b}^t + \mathbf{b}_0), \tag{5}$$

where $\boldsymbol{b}_0 \in \mathbb{R}^n$ represents the trainable baseline threshold, $\beta \in \mathbb{R}$ represents a constant scaling factor, $\boldsymbol{b}^t \in \mathbb{R}^n$ represents the state of the threshold at time $t$, $\boldsymbol{\rho} \in \mathbb{R}^n$ represents the constant decay of the threshold, $\mathbf{W}_a \in \mathbb{R}^{n \times m}$ and $\mathbf{H}_a \in \mathbb{R}^{n \times n}$ denote the trainable input and recurrent weight matrices influencing the threshold via axo-somatic synapses, where the axon of the presynaptic neuron is connected to the soma of the postsynaptic neuron. Figure 1c shows an illustration of the SNU-a and the lower right part of Fig. 1a shows an example of the membrane potential evolution.

Another type of neural connectivity in the human brain is via axo-axonic synapses, indicated with orange color in Fig. 1a. These synapses mediate the release of neurotransmitters from the presynaptic to the postsynaptic neuron. Interneurons (INs), for example Chandelier INs, that connect via axo-axonic synapses are found in the brain and often act as inhibitors for the postsynaptic neuron [30, 31]. However, it has been found that such neurons may be excitatory as well [31]. We incorporate the axo-axonic synapses into the SNU framework and propose SNU-o units, with the following equations

$$\boldsymbol{s}^t = \mathbf{g}(\mathbf{W}\boldsymbol{x}^t + \mathbf{H}\boldsymbol{y}^{t-1} + d \cdot \boldsymbol{s}^{t-1} \odot (\mathbb{1} - \tilde{\boldsymbol{y}}^{t-1})), \tag{6}$$

$$\tilde{\boldsymbol{y}}^t = \mathbf{h}(\boldsymbol{s}^t + \boldsymbol{b}^t), \tag{7}$$

$$\boldsymbol{y}^t = \tilde{\boldsymbol{y}}^t \odot \mathbf{o}\left(\mathbf{W}_o\boldsymbol{x}^t + \mathbf{H}_o\boldsymbol{y}^{t-1} + \boldsymbol{b}_o^t\right) \tag{8}$$

where $\tilde{\boldsymbol{y}}^{\boldsymbol{t}}$ represents the unmodulated output of the neuron driving the neural reset and $\boldsymbol{y}^t$ represents the modulated output of the neuron propagating to other connected neurons. $\mathbf{W}_o \in I\!R^{n \times m}$ and $\mathbf{H}_o \in I\!R^{n \times n}$ denote the trainable input and recurrent weight matrices and $\boldsymbol{b}_o^t \in I\!R^n$ denotes the trainable bias term. All three modulate the neuronal output via axo-axonic connections. Note that in our simulations we use the sigmoid function as the activation for the output modulation in (8), i.e., $\mathbf{o} = \boldsymbol{\sigma}$, to mimic the inhibitory character of these synapses. However, by using a different activation function $\mathbf{o}$, the outputs can also be modulated in an excitatory manner. Figure 1d shows an illustration of the SNU-o with the output modulation and the upper right part of Fig. 1a shows its connectivity motif along with an example of the membrane potential evolution.

## 3. RNN-T WITH BIOLOGICALLY-INSPIRED DYNAMICS

In this work, we redesign the RNN-T network architecture illustrated in Fig. 2 using the novel units introduced in Section 2. In order to do so, we follow a three-step approach. First, the units are introduced into the prediction network, replacing the LSTMs, while the encoder network remains composed of LSTM units. In a second step, we incorporate the sSNU variants into the encoder network only, i.e., the prediction network remains composed of LSTMs. Finally, the sSNU variants are integrated into both network parts and thus the full RNN-T architecture is composed of sSNU units. Since the encoder network and the prediction network carry out different tasks, different units might be better suited to be used in each of them. Such a composition of diverse unit and synapse types is also observed in the brain.

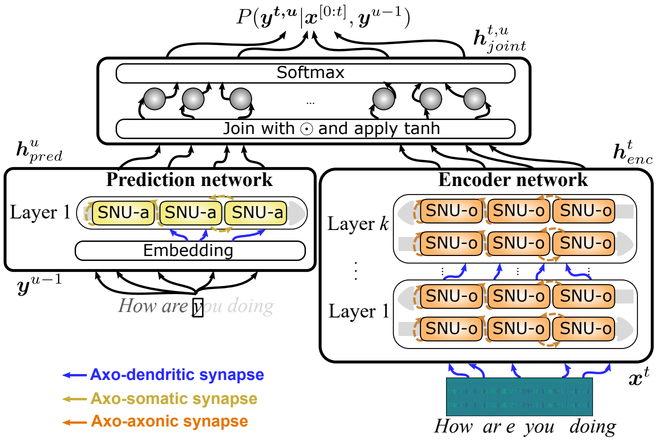

**Fig. 2**. **Illustration of the RNN transducer architecture**. An input vector $\boldsymbol{x}^t$, containing MFCC features of a speech signal, is processed by the encoder network, here represented with sSNU-o units. The prediction network, corresponding to a language model, here represented with sSNU-a units, enhances the predictions based on the past output sequence of the RNN-T, i.e. $\boldsymbol{y}^{u-1}$. The joint network forms the final predictions based on the outputs from both subnetworks.

ASR systems typically require a large amount of computing resources and thus reducing the computational cost is of paramount importance. Our proposed units not only reflect the biological inspirations, but additionally are simpler than LSTMs in terms of the number of gates and parameters. Therefore, the RNN-T network composed of these units provides the potential to drastically reduce the computational cost and latency.

## 4. RESULTS

The simulations were performed using the Switchboard speech corpus comprising roughly 300 hours of English telephone conversations between strangers on a pre-assigned topic. In order to be comparable to the state-of-the-art literature, we closely followed the data preprocessing, including the augmentation techniques, the evaluation, as well as the network architecture setup presented in [22]. In our simulations, an LSTM-based version of this RNN-T architecture achieves a word error rate (WER) of 12.7 % on the Hub5 2000 evaluation set, which we consider as our baseline. We investigated various configurations of the sSNU variants, which we abbreviated for simplicity. The trainable parameters of each variant are explicitly highlighted in Table 1.

The first part of Table 2 summarizes the results of the RNN-T architecture, where the sSNU variants were only integrated into the prediction network. Several configurations, e.g., sSNU *R* (12.4% WER), sSNU-a *R* (12.0% WER) and sSNU-o *R,Ro* (12.4% WER), outperform the LSTM-based variant (12.7% WER). Moreover, the number of trainable parameters as well as the number of multiplications of the prediction network, including vector-matrix and scalar multiplications, are substantially reduced.

In addition to the WER, we investigated the transcription time of the various models and we report the total time for the transcription of a single utterance. The last column in Table 2 shows the average time taken for the greedy decoding of an utterance with $T = 388$ input frames (∼7.8s of audio input). Note that the prediction network contributes only a small fraction to the total computational cost of the RNN-T network and thus the reduced number of multiplications has a negligible effect on the transcription time.

Next, the sSNU variants were integrated into the encoder network, while the prediction network remained composed of LSTMs. As one can see in the second part of Table 2, the number of trainable parameters and the number of multiplications of the encoder network are significantly reduced again. In fact, the reduction in the total number of parameters is much higher than that of an SNU-based prediction network. The sSNU-o *R,Ro* variant achieved 14.7% WER compared to 12.7% WER of the LSTM-based encoder network, but with a 50% reduction in the number of trainable parameters of the encoder network and a ∼40% reduction in the average time taken for decoding.

Finally, the sSNU variants were integrated into both subnetworks and hence the entire RNN-T was solely based on sSNUs. The last part of Table 2 summarizes the results of this scenario. Consistent with the prior two cases, the RNN-T architecture achieved competitive performance with 16.0% WER and 14.9% WER, but with a reduced number of parameters of the full RNN-T by ∼50% as well as a reduced latency by ∼40%.

Since latency is a critical metric for speech recognition systems, we investigated it further. Figure 3 depicts a comparison of the transcription time of utterances with various lengths. To ensure repro-

**Table 1**. sSNU acronyms used in the result tables and details on their trainable parameters.

| RNN *Suffix* | Comment | Threshold | Axo-dendritic | Axo-somatic | Axo-axonic |
|---|---|---|---|---|---|
| sSNU | Feedforward | $\mathbf{b}$ | $\mathbf{W}$ | | |
| sSNU *R* | Recurrent | $\mathbf{b}$ | $\mathbf{W}, \mathbf{H}$ | | |
| sSNU-a | Adaptive thr. feedforward | $\mathbf{b_o}$ | $\mathbf{W}$ | | |
| sSNU-a *R* | Adaptive thr. recurrent | $\mathbf{b_o}$ | $\mathbf{W}, \mathbf{H}$ | | |
| sSNU-a *R,Ra* | Adaptive thr. axo-somatic recurrent | $\mathbf{b_o}$ | $\mathbf{W}, \mathbf{H}$ | $\mathbf{H_a}$ | |
| sSNU-o | Output modulating feedforward | $\mathbf{b}$ | $\mathbf{W}$ | | $\mathbf{W_o}, \mathbf{b_o}$ |
| sSNU-o *R,Ro* | Output modulating recurrent | $\mathbf{b}$ | $\mathbf{W}, \mathbf{H}$ | | $\mathbf{W_o}, \mathbf{H_o}, \mathbf{b_o}$ |

ducibility and consistency, the examples were chosen such that the transcription output of all models was the same and the timing results were averaged over 10 repetitions. In general, the time taken to transcribe utterances increases proportionally to the utterance length. The RNN-T architecture composed of LSTM units, indicated with red dots, is the slowest among the evaluated models. Depending on the sSNU variant used, the inference time can be reduced significantly. For example, the RNN-T network using sSNU-o units for the encoder network and sSNU-o units for the prediction network (6×640 sSNU-o *R,Ro*, 1×768 sSNU-o *R,Ro*), has an approximately 40% reduced latency.

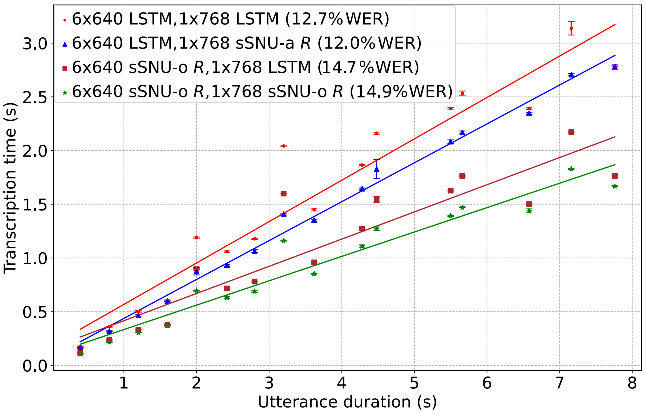

**Fig. 3**. **Comparison of the transcription time of various architectures.** The data points represent the mean time taken to decode utterances of different lengths. The markers indicate the various architectures used and the error bars depict the standard deviation over 10 executions.

## 5. CONCLUSIONS

In this work we combine insights from neuroscience with a state-of-the-art network architecture from machine learning and apply it to the task of speech recognition. Inspired by the diverse types of synapses present in the brain, we enhance the dynamics of the commonly used LIF neuron model with a threshold adaptation mechanism based on axo-somatic synapses as well as with an output modulating mechanism based on the axo-axonic synapses. We successively integrate these novel units into the RNN-T architecture and demonstrate that they enable end-to-end speech recognition with a competitive performance of 14.9% WER, compared to 12.7% WER for the LSTM-based RNN-T. Even more importantly, the introduced units are simpler than LSTMs, so that the computational costs as well as the latency can be reduced by 50% and 40% respectively. Finally, it is worth mentioning that biology provides an abundance of mechanisms which could further enrich the dynamics of neurons and that are not yet covered in commonly used neural network models. Our experimental results indicate that such mechanisms could potentially bolster the performance of biologically-inspired neural units and are therefore an essential step going forward.

**Table 2**. Performance comparison of the RNN-T network, where the sSNU variants are incorporated into different parts of the network.

| Prediction network | | | Encoder | WER | $t_i$(s) |
|---|---|---|---|---|---|
| RNN | # Par. | # Mult. | RNN | | |
| LSTM | 2.39M | 2.39M | LSTM | 12.7 | 2.78 |
| sSNU | 8.45k | 9.22k | | 15.1 | 2.71 |
| sSNU *R* | 0.60M | 0.60M | | 12.4 | 2.76 |
| sSNU-a | 8.45k | 11.52k | | 12.1 | 2.73 |
| sSNU-a *R* | 0.60M | 0.60M | | 12.0 | 2.78 |
| sSNU-o | 16.90k | 17.66k | | 12.4 | 2.75 |
| sSNU-o *R,Ro* | 1.20M | 1.20M | | 12.6 | 2.76 |

| Pred. | Encoder | | | WER | $t_i$(s) |
|---|---|---|---|---|---|
| RNN | RNN | # Par. | # Mult. | | |
| LSTM | LSTM | 54.20M | 54.20M | 12.7 | 2.78 |
| | sSNU-a *R,Ra* | 18.47M | 18.50M | 25.2 | 1.23 |
| | sSNU-o | 17.27M | 17.28M | 23.2 | 1.22 |
| | sSNU-o *R,Ro* | 27.10M | 27.11M | 14.7 | 1.76 |

| Pred. RNN | Enc. RNN | # Par. | # Mult. | WER | $t_i$(s) |
|---|---|---|---|---|---|
| LSTM | LSTM | 56.59M | 56.58M | 12.7 | 2.78 |
| sSNU-a *R* | sSNU-o *R,Ro* | 27.70M | 27.71M | 16.0 | 1.64 |
| sSNU-o *R,Ro* | sSNU-o *R,Ro* | 28.30M | 28.30M | 14.9 | 1.66 |

## 6. REFERENCES

[1] Alex Graves, Douglas Eck, Nicole Beringer, and Juergen Schmidhuber, "Biologically plausible speech recognition with LSTM neural nets," in *Biologically Inspired Approaches to Advanced Information Technology*, Auke Jan Ijspeert, Masayuki Murata, and Naoki Wakamiya, Eds., Berlin, Heidelberg, 2004, pp. 127–136, Springer Berlin Heidelberg.

[2] Biing-Hwang Juang and Lawrence R Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, vol. 1, pp. 67, 2005.

[3] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The Microsoft 2016 conversational speech recognition system," *ICASSP*, pp. 5255–5259, Mar 2017.

[4] Thai-Son Nguyen, Sebastian Stueker, and Alex Waibel, "Super-Human Performance in Online Low-latency Recognition of Conversational Speech," *arXiv*, Oct 2020.

[5] Dong Wang, Xiaodong Wang, and Shaohe Lv, "An Overview of End-to-End Automatic Speech Recognition," *Symmetry*, vol. 11, no. 8, pp. 1018, Aug 2019.

[6] G. Rao, "A Comparison of End-to-End Speech Recognition Architectures in 2021," January 2021, https://www.assemblyai.com/blog/a-survey-on-end-to-end-speech-recognition-architectures-in-2021.

[7] Alex Graves, "Sequence Transduction with Recurrent Neural Networks," *arXiv*, Nov 2012.

[8] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016, pp. 4960–4964.

[9] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Interspeech*, 2020, pp. 5036–5040.

[10] Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo-yiin Chang, Kanishka Rao, and Alexander Gruenstein, "Streaming End-to-end Speech Recognition For Mobile Devices," *arXiv*, Nov 2018.

[11] Tara Sainath, Ruoming Pang, David Rybach, Yanzhang He, Rohit Prabhavalkar, Wei Li, Mirkó Visontai, Qiao Liang, Trevor Strohman, Yonghui Wu, Ian McGraw, and Chung-Cheng Chiu, "Two-pass end-to-end speech recognition," 09 2019, pp. 2773–2777.

[12] Ke Hu, Tara N. Sainath, Ruoming Pang, and Rohit Prabhavalkar, "Deliberation Model Based Two-Pass End-to-End Speech Recognition," *arXiv*, Mar 2020.

[13] Ke Hu, Ruoming Pang, Tara N. Sainath, and Trevor Strohman, "Transformer based deliberation for two-pass speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 68–74.

[14] Jonathan Dennis, Qiang Yu, Huajin Tang, Huy Dat Tran, and Haizhou Li, "Temporal coding of local spectrogram features for robust sound recognition," in *ICASSP*, 2013, pp. 803–807.

[15] Jibin Wu, Yansong Chua, and Haizhou Li, "A biologically plausible speech recognition framework based on spiking neural networks," in *IJCNN*, 2018, pp. 1–8.

[16] Linhao Dong and Bo Xu, "CIF: Continuous Integrate-and-Fire for End-to-End Speech Recognition," *arXiv*, May 2019.

[17] Robert Gütig and Haim Sompolinsky, "The tempotron: a neuron that learns spike timing–based decisions - Nature Neuroscience," *Nat. Neurosci.*, vol. 9, no. 3, pp. 420–428, Mar 2006.

[18] Cong Shi, Tengxiao Wang, Junxian He, Jianghao Zhang, Liyuan Liu, and Nanjian Wu, "DeepTempo: A Hardware-Friendly Direct Feedback Alignment Multi-Layer Tempotron Learning Rule for Deep Spiking Neural Networks," *IEEE Trans. Circuits Syst. II*, vol. 68, no. 5, pp. 1581–1585, Mar 2021.

[19] Stanisław Woźniak, Angeliki Pantazi, Thomas Bohnstingl, and Evangelos Eleftheriou, "Deep learning incorporating biologically inspired neural dynamics and in-memory computing," *Nat. Mach. Intell.*, vol. 2, pp. 325–336, Jun 2020.

[20] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," *arXiv*, Dec 2014.

[21] Ilya Loshchilov and Frank Hutter, "Decoupled Weight Decay Regularization," *arXiv*, Nov 2017.

[22] George Saon, Zoltan Tueske, Daniel Bolanos, and Brian Kingsbury, "Advancing RNN Transducer Technology for Speech Recognition," *arXiv*, Mar 2021.

[23] Allyson Howard, Gabor Tamas, and Ivan Soltesz, "Lighting the chandelier: new vistas for axo-axonic cells," *Trends Neurosci.*, vol. 28, no. 6, pp. 310–316, Jun 2005.

[24] Mark Bear, Barry Connors, and Michael A Paradiso, *Neuroscience: Exploring the brain*, Jones & Bartlett Learning, LLC, 2020.

[25] Allen Institute for Brain Science, "Allen Human Brain Atlas," 2010.

[26] Stanisław Woźniak, Angeliki Pantazi, Thomas Bohnstingl, and Evangelos Eleftheriou, "Deep learning incorporating biologically-inspired neural dynamics," *arXiv*, Dec 2018.

[27] Guillaume Bellec, Darjan Salaj, Anand Subramoney, Robert Legenstein, and Wolfgang Maass, "Long short-term memory and Learning-to-learn in networks of spiking neurons," in *NeurIPS*, 2018, pp. 787–797.

[28] Bertrand Fontaine, José Luis Peña, and Romain Brette, "Spike-Threshold Adaptation Predicted by Membrane Potential Dynamics In Vivo," *PLoS Comput. Biol.*, vol. 10, no. 4, pp. e1003560, Apr 2014.

[29] Chao Huang, Andrey Resnik, Tansu Celikel, and Bernhard Englitz, "Adaptive Spike Threshold Enables Robust and Temporally Precise Neuronal Encoding," *PLoS Comput. Biol.*, vol. 12, no. 6, pp. e1004984, Jun 2016.

[30] Robin Tremblay, Soohyun Lee, and Bernardo Rudy, "Gabaergic interneurons in the neocortex: from cellular properties to circuits," *Neuron*, vol. 91, no. 2, pp. 260–292, 2016.

[31] Gord Fishell and Bernardo Rudy, "Mechanisms of Inhibition within the Telencephalon: "Where the Wild Things Are"," *Annu. Rev. Neurosci.*, vol. 34, no. 1, pp. 535–567, Jun 2011.