

# Hardware-Friendly Synaptic Orders and Timescales in Liquid State Machines for Speech Classification

Vivek Saraswat, Ajinkya Gorad, Anand Naik, Aakash Patil and Udayan Ganguly

Dept. of Electrical Engineering  
Indian Institute of Technology, Bombay, Mumbai, India  
[udayan@ee.iitb.ac.in](mailto:udayan@ee.iitb.ac.in)

**Abstract**—Liquid State Machines are brain inspired spiking neural networks (SNNs) with random reservoir connectivity and bio-mimetic neuronal and synaptic models. Reservoir computing networks are proposed as an alternative to deep neural networks to solve temporal classification problems. Previous studies suggest 2<sup>nd</sup> order (double exponential) synaptic waveform to be crucial for achieving high accuracy for TI-46 spoken digits recognition. The proposal of long-time range (ms) bio-mimetic synaptic waveforms is a challenge to compact and power efficient neuromorphic hardware. In this work, we analyze the role of synaptic orders namely:  $\delta$  (high output for single time step), 0<sup>th</sup> (rectangular with a finite pulse width), 1<sup>st</sup> (exponential fall) and 2<sup>nd</sup> order (exponential rise and fall) and synaptic timescales on the reservoir output response and on the TI-46 spoken digits classification accuracy under a more comprehensive parameter sweep. We find the optimal operating point to be correlated to an optimal range of spiking activity in the reservoir. Further, the proposed 0<sup>th</sup> order synapses perform at par with the biologically plausible 2<sup>nd</sup> order synapses. This is substantial relaxation for circuit designers as synapses are the most abundant components in an in-memory implementation for SNNs. The circuit benefits for both analog and mixed-signal realizations of 0<sup>th</sup> order synapse are highlighted demonstrating 2-3 orders of savings in area and power consumptions by eliminating Op-Amps and Digital to Analog Converter circuits. This has major implications on a complete neural network implementation with focus on peripheral limitations and algorithmic simplifications to overcome them.

**Keywords**—LSM, reservoir, speech classification, SNNs, synapse, order, timescale

## I. INTRODUCTION

Spiking Neural Networks (SNNs) are the third generation of artificial neural networks [1]. Information is encoded in the timing of spikes of the neurons which results in temporally-specific and low-power communication events in the neural network [2]. SNNs have been adopted for a large number of classification and pattern recognition tasks with a focus on developing dedicated neuromorphic hardware to harness the power efficiency of SNNs [3], [4]. Although, biologically more plausible, SNNs face significant challenges in mapping learning algorithms to spiking neurons [5], [6] as well as in circuit design [7]. Liquid State Machines (LSMs) are an attempt to simplify the network development and learning strategies by borrowing further inspiration from the human cortex. LSM is a

spiking neural network architecture which has a reservoir of recurrently connected neurons [8]. This architecture is proposed as an alternative to Deep Neural Networks which comprise of a large number of successive layers of neurons. These layers can differ in size (number of neurons per layer) and the objective is to have sufficient number of learning parameters (inter-layer weights and neuronal biases) so as to achieve arbitrary classification functionality. LSMs, on the other hand, deviate from this depth of network idea. In the reservoir computing framework, the presence of recurrent dynamics and higher dimensionality of information represented in the reservoir is entrusted with facilitating arbitrary separation and generalization functions necessary for classification tasks. LSMs have been shown to be particularly well suited for temporal information datasets like speech and video activity recognition [9]–[12]. The reservoir response is expected to act as a universal function generator and all learning is pushed to a single linear classifier outside the reservoir itself [8].

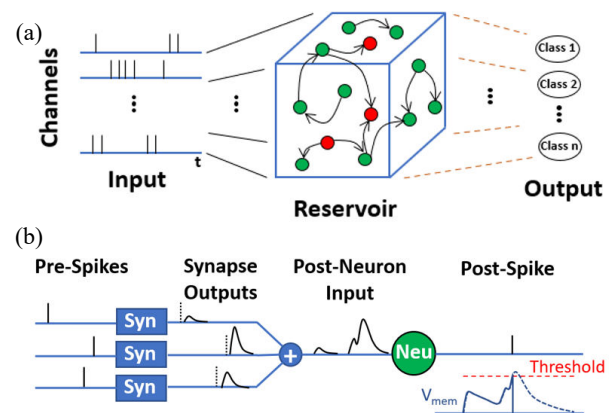


Fig. 1. (a) LSM Architecture – spiking input channels, randomly connected reservoir of excitatory (green) and inhibitory (red) neurons and output classifier neurons, the input and reservoir weights are fixed while the classifier weights are trained, (b) Basic computations in a reservoir of leaky-integrate and fire (LIF) neurons: Pre-spikes are scaled and shaped by synapses, then summed and integrated by the post-neuron, a spike is issued when membrane potential reaches the threshold and the potential is reset

Figure 1 shows an example LSM. There is an input layer of spiking neurons (called input channels), a large reservoir of recurrently connected neurons and an output layer (Fig. 1(a)). The input spikes are generated from the raw input features using a preprocessing step which is specific to the dataset. The reservoir connectivity follows cortical brain-inspired random

This work has been supported by Prime Minister's Research Fellowship, Ministry of Electronics and Information Technology, Govt. of India. Ajinkya Gorad is currently affiliated with Aalto University, Finland.

connectivity. Different models exist like the local probabilistic connections, axonal model and small-world networks [13]. Once generated using a random model, the reservoir connections are not changed during the network operation. A single linear classifier between the reservoir neurons and the output neurons is chosen to be trained. The basic computations that happen in a recurrent network are shown in Fig. 1(b). The spikes issued by different pre-neurons arrive at synapses of these neurons with the post neuron. Each synapse outputs a post-synaptic current in response to the pre-neuron spike after a small fixed delay. The synapses output a scaled current waveform. The amplitude depends on the connection strength of the synapse. The summed current is then integrated by the post-neuron which is typically modelled as a leaky integrate and fire neuron. An output spike is issued by the post neuron if the integrated membrane potential reaches a threshold and the potential is reset to resting value.

A recurrent neural network (like the reservoir) is best implemented as an in-memory architecture where neurons are able to interact with each other by means of the more numerous synaptic connections in a parallel manner [14]–[16]. The model proposed for spiking neurons is typically the leaky-integrate and fire neurons [17]. There are different models of the post-synaptic waveforms of the unit strength synapse in response to an input spike (Fig. 2): (a)  $\delta$  synapse outputs high current for a single time step, (b) 1<sup>st</sup> order synapse outputs an exponential decaying current and (c) 2<sup>nd</sup> order synapse outputs double exponential rise and fall waveform. Typically, the biomimetic 2<sup>nd</sup> order synapse waveform is used by the biologically inspired algorithms [18]. Recently, a digital implementation of an LSM using these models was proposed for the TI-46 spoken digits recognition task with a spike based local learning rule for the linear classifier [17]. Further, the LSM network was represented using a state-space model and a performance predicting memory metric was extracted [11]. It has been argued that the post synaptic current waveform has a crucial role to play in the classification accuracy [18]–[20] specifically for the speech digit recognition task [17] (Table I). Error was shown to degrade by more than 10 times if a  $\delta$  synapse is employed in place of the 2<sup>nd</sup> order synapse waveform keeping the connection strengths unchanged [17]. This immediately renders the circuit-friendly  $\delta$  synapse unfeasible for circuit designers implementing SNNs. Complex waveforms of higher (1<sup>st</sup> and 2<sup>nd</sup>) order synapses are not circuit-friendly [21]. The area and power consumed by circuits realizing long-time (ms) range waveforms especially relevant for real-time sensory input data is a challenge for neuromorphic chip designers [22].

Algorithmic simplifications to waveforms requirement can have significant impact for hardware realizations of SNNs. Hence, in this paper, first we show the strategy to obtain the optimal operating point and its relation with reservoir spiking activity for a given unit synapse waveform and timescale. For this we perform a more comprehensive parameter sweep on the connection strengths for the TI-46 spoken digits recognition task. Next, we propose a 0<sup>th</sup> order synapse (Fig. 2(d)) with a

timescale which is essentially a rectangular pulse of finite width to get the best of both performance and circuit friendly implementation.  $\delta$  synapse is a special case of the 0<sup>th</sup> order synapse where pulse width equals a single time step. Finally, we perform a circuit cost and benefits estimation for the different synaptic orders that greatly affects the circuit design choices by the SNN chip designers.

TABLE I. PERFORMANCE OF DIFFERENT SYNAPSE ORDERS [17]

Metric	$\delta$ Synapse	1 <sup>st</sup> Order Synapse	2 <sup>nd</sup> Order Synapse
Accuracy (%)	88.85	90.73	99.09
Error (%)	11.15	9.27	0.91
Feature	Circuit-friendly	Biomimetic	Biomimetic

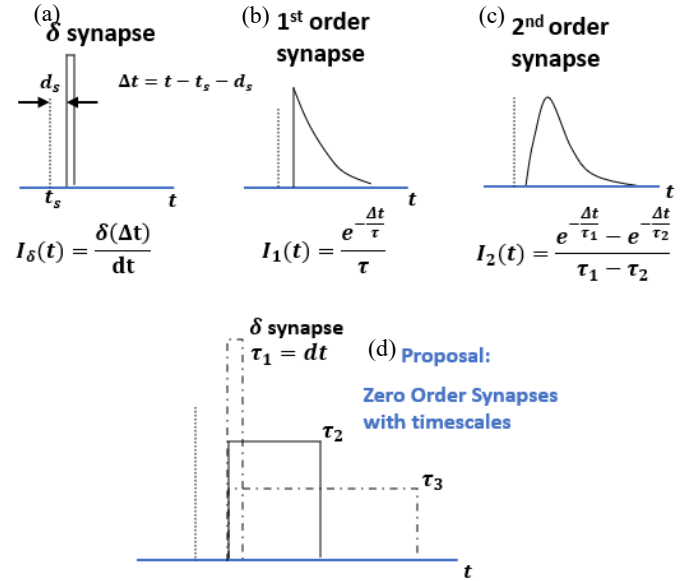


Fig. 2. Different types of unit synapses: (a)  $\delta$  (b) 1<sup>st</sup> order (c) 2<sup>nd</sup> order synapses. Dotted line is the spike time,  $t_s$ , in response to which the synaptic current waveform is output after a delay,  $d_s$ . (d) Proposed 0<sup>th</sup> order rectangular synapse with finite pulse width. A synapse can have different timescales ( $\tau_1, \tau_2, \tau_3$ ) but the total integrated current is assumed constant for a unit synapse. A  $\delta$  synapse is a special case of 0<sup>th</sup> order synapse where pulse width is equal to a single time-step  $dt$ .

## II. TI-46 SPOKEN DIGITS RECOGNITION SETUP

The TI-46 spoken digits dataset comprises of 5 speakers uttering 10 times each of the 10 digits (500 samples) [17]. Each utterance is about 1 second long. The preprocessing of the audio waveforms to achieve a spiking input for the LSM is well established and follows the Lyon's Passive Ear Model [23]. This is a human ear Cochlea inspired model which works with 77 channels of band-pass filters followed by Automatic Gain Control and resampling stages. Finally, Bens Spiking Algorithm employs a rate coding-based scheme to output spike trains in 77

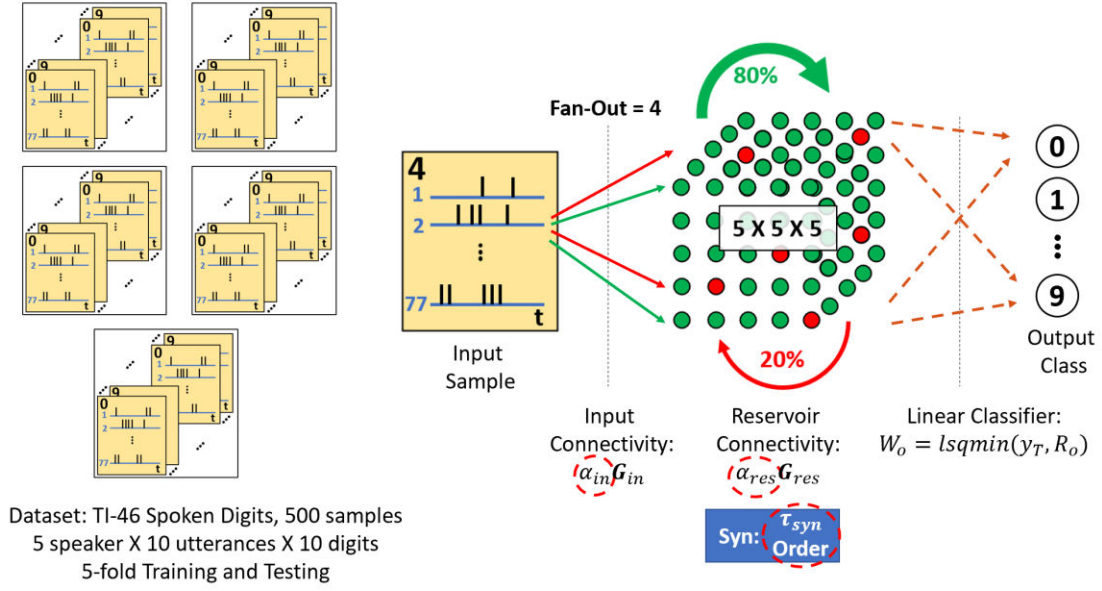


Fig. 3. TI-46 preprocessed dataset and spoken digit recognition setup. The quantities encircled by red dashed lines are the design space parameters for this study.

channels for each input digit sample. The details of the preprocessing stage have been discussed in detail elsewhere [11], [24]. These 500 samples are used for training and testing the network in a 5-fold manner. 400 samples are used to train the network and remaining 100 are tested in a rotating manner (Fig. 3).

The reservoir consists of 125 neurons. The reservoir connectivity ( $\mathbf{G}_{res}$ ) is established using the local probabilistic random connections model with 80% excitatory neurons and 20% inhibitory neurons [11]. This model requires the neurons to be arranged on a grid (5 X 5 X 5 chosen here) to calculate the connection probability as a function of the distance between the neurons in the grid structure. The exact parameters for reservoir generation are discussed in detail in [11]. Each spike train channel of any input digit sample is connected to 4 randomly chosen reservoir neurons with equal number of excitatory and inhibitory connections of equal strength. This results in an input connectivity matrix ( $\mathbf{G}_{in}$ ) (Fig. 2). Although, the connectivity matrices  $\mathbf{G}_{in}$  and  $\mathbf{G}_{res}$  are not changed, a constant scaling factor to synaptic strengths  $\alpha_{in}$  and  $\alpha_{res}$  are introduced as tuning parameters. These decide the operating point of the network for any given synapse model. The reservoir output spikes feed into the output neurons using the classifier weights. The reservoir output spikes for all neurons are averaged over the entire duration of the sample to get the reservoir output response for training. This reservoir response for all training samples and the target output layer response is used to train the classifier weights by a least squares minimization algorithm:

$$W_o = \min_w ||Ky_T - WR_o||^2 \quad (1)$$

where  $W_o$  are the optimal output weights (artificially limited to  $\pm W_{lim}$ ),  $K$  is a scaling constant,  $y_T$  is the expected target spiking in the output layer collated for all training samples and  $R_o$  is the time-averaged reservoir output response. The trained weights are then tested on the testing samples to calculate the speech classification accuracy by identifying the maximum spiking activity output neuron.

The neurons are modelled as identical Leaky Integrate and Fire neurons. The synaptic order and timescale are also identical for all synapses in the network. However, we have the option of choosing the synapse order and timescale of the unit synapse apart from the previously mentioned input and reservoir weights scaling,  $\alpha_{in}$  and  $\alpha_{res}$ , in this simulation setup (Table II).

TABLE II. SIMULATION SETUP

Component	Parameter	Values
Preprocessing [Zhang et al.] [17]	Model	Lyon's Passive Cochlea
	Model	Local Probabilistic (80E/20I)
Reservoir [Gorad et al.] [11]	$\alpha_{in}$	1 – 20
	$\alpha_{res}$	0.1 – 4
Synapse	Order	$\delta$ , 0, 1, 2
	Timescale $\tau$	1 – 50 ms
Neuron	Leakage $\tau_N$	64 ms
	Refractory Period	2 ms
	Threshold	20 mV
	Resting	0 mV
Classifier	Model	Least Squares
	$W_{lim}$	8
	$K$	1000



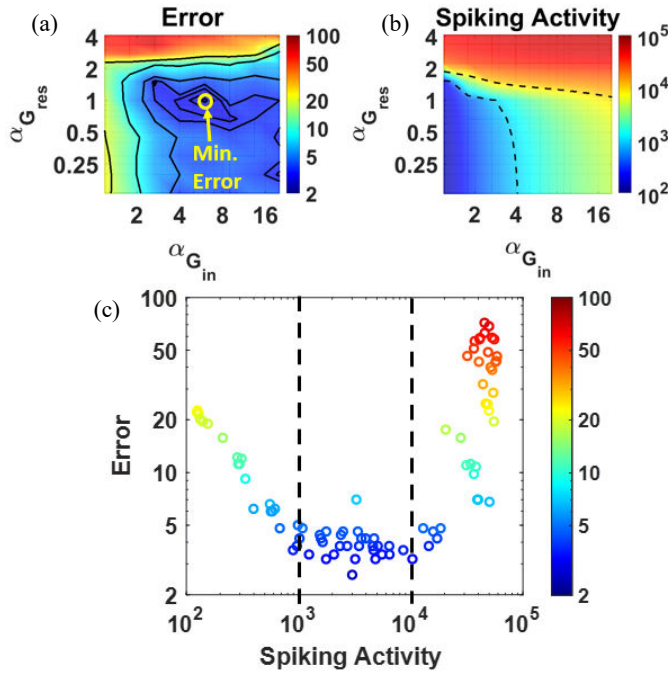


Fig. 4. Experiments for 2<sup>nd</sup> order synapse with timescales  $(\tau_1, \tau_2) = (8 \text{ ms}, 4 \text{ ms})$  – the fall and rise timescales. (a) Error as a function of the input and reservoir weights scaling – minimum error is highlighted and equi-error contours are plotted, (b) Spiking activity in the reservoir averaged over all the input samples as a function of the input and reservoir weights scaling, (c) Error as a function of the spiking activity in the reservoir shows an optimal range of reservoir spiking activity (denoted by dashed lines)

### III. RESULTS AND DISCUSSION

#### A. Effect of weights scaling

In order to evaluate the peak performance of a given synaptic order and timescale, we perform a parameter space sweep for  $\alpha_{in}$  and  $\alpha_{res}$ . There is an optimal  $\alpha_{in}$  and  $\alpha_{res}$  point where the 5-fold average classification accuracy is maximum or the error is minimum (Fig. 4(a)). The weight scaling parameters jointly control the level of spiking activity in the reservoir, i.e., spikes issued by all reservoir neurons averaged over all the input samples. As expected, the reservoir spiking activity rises monotonically with  $\alpha_{in}$  and  $\alpha_{res}$  (Fig. 4(b)). The response w.r.t  $\alpha_{res}$  is more non-linear due to the recurrent nature of the reservoir connections setting up a positive feedback and chaotic dynamics [11]. A more insightful representation is to observe the variation of the error with the spiking activity of the reservoir (which takes into account the net effect of both the weights scaling). Error is minimum for an optimal range of spiking activity (Fig. 4(c)). The error rises rapidly above and below this optimal spiking activity range. Thus, the peak performance for a given synaptic order and timescale is evaluated from its error vs spiking activity graph. This is aligned with the well-known “edge of chaos” theory for maximum performance [11]. The observed optimal spiking activity range depends on the actual application and the associated network topology. Nonetheless, a non-monotonic trend in the achieved error rates is expected with the spiking activity in general. Hence, this framework provides a method to bias the reservoir in a high-performing region. Although, the optimal operating point has been achieved by tuning the weights scaling, ultimately, the effect of this tuning is

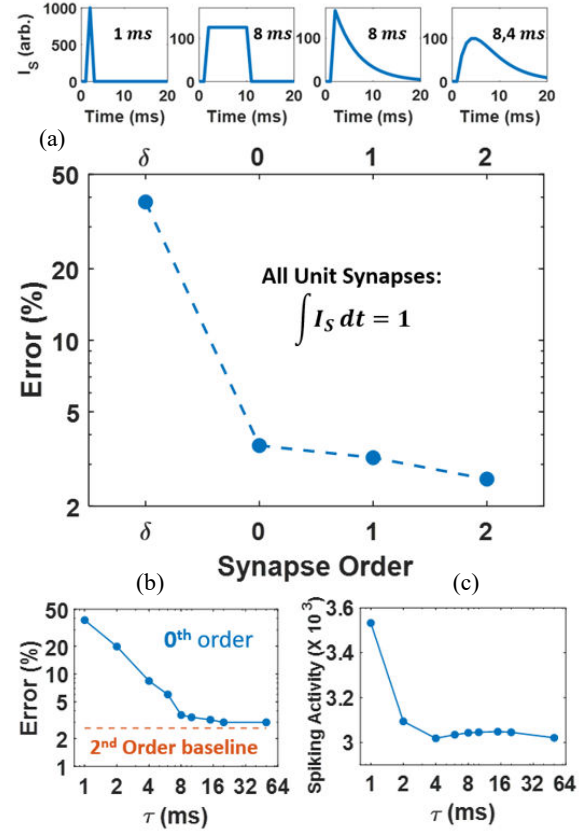


Fig. 5. Fixed  $(\alpha_{in}, \alpha_{res}) = (6, 1)$  simulations for (a) Error as a function of the synapse order; synaptic timescales of different orders are as shown in the inset at the top (synapses follows constant integrated current model when order or timescale is varied), (b) Error as a function of the synaptic timescale of the 0<sup>th</sup> order synapse and comparison with the 2<sup>nd</sup> order  $(\tau_1, \tau_2) = (8 \text{ ms}, 4 \text{ ms})$  baseline, (c) Spiking activity in the reservoir as a function of the synaptic timescale of the 0<sup>th</sup> order synapse

a modulation of the reservoir spiking activity. In practice, this may be achieved not just by synaptic plasticity but also by altering neuronal excitability and/or timescales. Efficient hardware implementations of controllable synaptic input integration neurons [25], [26] and learning synapses [27], [28] is key and can benefit LSMs hugely.

#### B. Effect of synapse order and timescale

The optimal  $\alpha_{in}$  and  $\alpha_{res}$  determined from the 2<sup>nd</sup> order synapse experiments are now used for testing the other unit synaptic orders. Whereas the  $\delta$  synapse shows a marked deterioration in error, the 0<sup>th</sup> and the 1<sup>st</sup> order synapses are still comparable to 2<sup>nd</sup> order performance provided they have comparable timescales (Fig. 5(a)). This is significant since 0<sup>th</sup> order synapses are much more circuit-friendly than higher order synapses as discussed in Section IV. Next, we perform a synapse timescale variation study for the 0<sup>th</sup> order synapse for the fixed  $\alpha_{in}$  and  $\alpha_{res}$  (Fig. 5(b)).  $\delta$  synapse is a special case of the 0<sup>th</sup> order synapse where pulse width equals a single time step (1 ms). When the synaptic timescale is varied, the integrated current is kept unchanged (constant charge or unit synapse variation). This leads to smaller timescales affecting the membrane potentials of the leaky integrating neurons more significantly and a sharp rise in the spiking activity of the reservoir for the  $\delta$  synapse (Fig. 5(c)). As identified earlier this

increased spiking activity is accompanied by a reduced classification accuracy for the  $\delta$  synapse. However, when the synaptic timescale of the 0<sup>th</sup> order synapse is comparable to the optimal 2<sup>nd</sup> order synapse timescale ( $\sim 8$  ms), the accuracies achieved by both the synapses are very similar (Fig. 5(b)). Hence, the 0<sup>th</sup> order synapse is as high performing as the second order synapse for the inference tasks if the timescale is allowed to be tuned.

### C. Peak Performance of all synaptic orders and timescales

We observe that the previous prediction of  $\delta$  synapse with fixed  $\alpha_{in}$  and  $\alpha_{res}$  being unsuitable for the speech classification LSM is a limited experiment (Table I and Fig. 5(a)). Hence, we

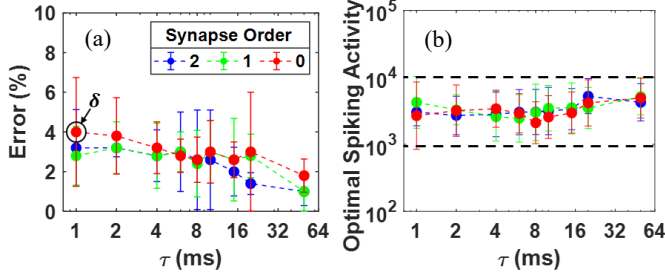


Fig. 6. (a) Minimum Error as a function of the synapse timescale for different synaptic orders calculated by sweeping weights scaling parameters to obtain the optimal operating point, error bars are plotted using the standard deviation of the 5-fold testing data per simulation. (b) Optimal reservoir spiking activity (mean and error bars) for the 10 lowest Error operating points during weights scaling sweep as a function of the synapse timescale for different synaptic orders validating the optimal spiking range framework.

TABLE III. COMPARISON OF TI-46 SPOKEN DIGIT CLASSIFICATION

Work	Synapse Order	Operating Point	Accuracy (%)
Verstraeten et al. [24]	Biomimetic 2 <sup>nd</sup> order	-	98
Zhang et al. [17]	Circuit-friendly $\delta$	Does not change with synapse order and $\tau$	89
	Biomimetic 1 <sup>st</sup> order		91
	Biomimetic 2 <sup>nd</sup> order		99
This Work	Circuit-friendly $\delta$	Optimal reservoir spiking activity strategy	96
	Circuit-friendly 0 <sup>th</sup> order		98
	Biomimetic 1 <sup>st</sup> order		99
	Biomimetic 2 <sup>nd</sup> order		99

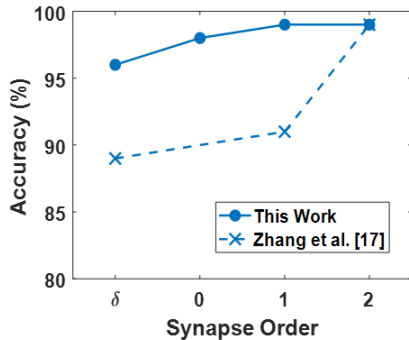


Fig. 7. Optimal classification accuracy as a function of the synapse order for the reservoir spiking activity based optimal weights scaling operating point proposed in this work compared to the fixed operating point results of previous works

next calculate the peak performance of different synaptic orders and timescales according to the strategy highlighted in Section III.A. We allow a scan over the  $\alpha_{in}$  and  $\alpha_{res}$  parameters to identify the best average 5-fold accuracy for a given unit synaptic waveform and then repeat this for 0<sup>th</sup>, 1<sup>st</sup> and 2<sup>nd</sup> order synapses of different timescales (Fig. 6(a)). For 2<sup>nd</sup> order, the  $\tau$  plotted is the fall  $\tau$  with the rise  $\tau$  being one-half of the fall  $\tau$ . A more comprehensive method of obtaining optimal operating point demonstrates that even  $\delta$  synapses are not much worse off compared to the optimal timescale 2<sup>nd</sup> order synapse (Table III). The peak performances for all orders and timescales are correlated to an optimal range of reservoir spiking activity (Fig. 6(b)). Thus, the optimal operating point strategy shows the true performance utility of different synapse models in contrast to previous studies (Fig. 7). This opens an avenue for the circuit designers to opt for 0<sup>th</sup> order synapses as a realistic choice for hardware implementation of such biologically inspired algorithms. The 0<sup>th</sup> order synapse have numerous circuit implementations related benefits which we discuss next.

## IV. CIRCUIT COST ESTIMATION

Spiking Neural Networks are a fundamentally different computing architecture with integrated memory and computational units and temporal information encoding. Whereas the conventional von-Neumann microprocessors and memories scale in terms of speed and area, they are primarily serial processors with each of the components (ALUs, cache) speeding up in time to improve overall throughput and efficiency.

Neuromorphic chips, which implement SNNs, tend to focus on radically different aspects. The computation is massively parallel and often an in-memory approach is used to accelerate the temporal vector-matrix multiplication [14]–[16]. The neuronal outputs are scaled and summed to generate the next stage inputs in a parallel manner. The weight matrix or the connection matrix becomes the most dominant circuit. An efficient connection unit or the synapse, hence, plays a critical role in determining how much power and area consuming, the SNN chip is going to be. This becomes a circuit design issue compounded by the fact that many real-time temporal classification tasks like speech processing require the implementation of long timescales (in the range of milliseconds) for the neuron's leakage and synaptic waveforms. Longer timescales are linked to increased design capacitance and hence lead to higher circuit area and power consumption [22]. This issue is typically overcome by operating the CMOS circuits in subthreshold regimes or making use of novel physics like impact ionization or band-to-band tunneling based neurons and synapses [25]–[27], [29], [30]. Both these approaches are under extensive investigation and show immense promise however they are plagued with variability and latency issues [29], [31]. It is possible that the SNN based algorithms may be robust to such component level variabilities if they employ a feedback mechanism [32]. Nonetheless, circuit-friendly simplifications to biologically motivated algorithms can go a long way in allowing conventional CMOS circuits to be well suited for designing neuromorphic chips.

One such simplification is proposed in this paper: the synapse order. Previously proposed algorithms for spoken digits classification have claimed a 2<sup>nd</sup> order synapse to be crucial for performance [17]. However, as shown in Section III.C, for any choice of synapse order and timescale, the optimal operating point requires a design space exploration on weights scaling to achieve optimal reservoir spiking activity. As a result, the degradation in performance by choosing a 0<sup>th</sup> order synapse is demonstrated to be much lower compared to previous claims (Table III and Fig. 7). Furthermore, the absolute accuracy for the  $\delta$  synapse is also high (96 %). The accuracy further rises ( $\sim 98$  %) for the 0<sup>th</sup> order synapses if the timescale of the rectangular pulse is allowed to increase and become comparable to 2<sup>nd</sup> order pulse timescales (Fig. 6(a)). This knowledge addition drastically impacts the choices of circuit designers now. The 0<sup>th</sup> order synapses are feasible and provide excellent performance. Further, the manifold circuit benefits that accompany circuit implementation of the 0<sup>th</sup> order synapse compared to 2<sup>nd</sup> order are presented next:

#### A. Benefits for analog implementation

Analog waveform shaping involves charging/discharging large capacitors followed by signal buffering circuits as drivers [22], [27], [33]. For e.g. to generate timescales in the range of milliseconds using minimum sized 45 nm CMOS technology transistors biased in the subthreshold regime requires a capacitance of 150 fF which is about 100 times the gate capacitance of the minimum sized transistor in that technology [34], [35]. This indicates that the majority of the synapse area will be occupied by the waveform shaping circuitry primarily the capacitance. As the number of timescales to be implemented in the waveform increase (as in higher order synapses), so do the number of capacitances. It is possible to achieve second order waveforms using single capacitances by differential pair integrator (DPI) circuits [19], [36] (Table IV). However, they have increased circuit complexity and bias generation circuits requirement rendering them vulnerable to device variability. Also, any analog signal needs to be buffered through a driver circuit to drive the next stages in the network without loading itself. A general analog waveform can be buffered using an operational amplifier which is a big ( $\sim 1000 \mu\text{m}^2$  in 0.25  $\mu\text{m}$  CMOS technology node) and a complex circuit affected by static bias power consumption ( $\sim 100\text{-}500 \mu\text{W}$ ) and variability [37], [38] (Table V). A binary level digital signal like a 0<sup>th</sup> order synapse, on the other hand, can be buffered much easily by a small ( $\sim 10 \mu\text{m}^2$  in 0.25  $\mu\text{m}$  node) inverter pair in series eliminating the need of a more complex drive circuitry. Since digital buffers only consume dynamic power with minimal static leakage, the power consumption ( $\sim C_{\text{load}} V_{\text{DD}}^2 / \tau$ ) at 1 ms timescale range is greatly reduced ( $\sim 100 \text{ nW}$  for 4 pF  $C_{\text{load}}$  and 3.3 V supply) compared to the analog buffer [39].

In addition, any rise and fall time of an analog waveform is determined by both the charging/discharging timescale and the location of the steady-state signal value w.r.t the present signal value. This means that constraints on the rise and fall times (as in higher order synapses) require the steady state voltage or the  $V_{\text{DD}}$  to be farther away from the unit synapse waveform amplitude. This introduces a voltage margin requirement. Contrary to this, on/off pulses for 0<sup>th</sup> order synapses without specific constraints on charging/discharging rates can charge all

the way to  $V_{\text{DD}}$  and GND without the requirement of a voltage margin. Thus, analog circuit design for 0<sup>th</sup> order synapses has significant benefits compared to higher order synapses.

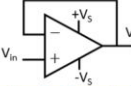

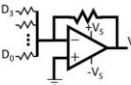
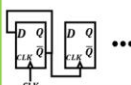
#### B. Benefits for mixed-signal implementation

In order to implement a waveform digitally, the waveform needs to be quantized at some levels (say 4-bits for sufficient precision). These levels are stored in a look-up table (LUT) which is a memory bank. In-memory multiply and accumulate operation depend on the synapse outputs to be analog currents that can be summed together in a parallel manner regardless of whether the neurons or synapses themselves are digital or analog implementations. Hence, at the time of application of the synapse output waveform, a digital-to-analog converter (DAC) is required (Table IV). The DAC is a complex mixed signal circuit whose power ( $500 \mu\text{W}$  for 4-bit and 0.25  $\mu\text{m}$  node) and area ( $\sim 1500 \mu\text{m}^2$ ) scale badly with number of bits as may be the requirement for a high-fidelity reconstruction of the desired waveform [40]–[42] (Table V). A 0<sup>th</sup> order synapse or a rectangular pulse of finite width, on the other hand, is much easier to implement using binary levels digital counters comprising of simple flip-flops ( $\sim 20$  transistors per flip-flop) with minimal circuit complexity ( $500 \mu\text{m}^2$  for 4-bit i.e., around 80-100 transistors in 0.25  $\mu\text{m}$  node) and power consumption ( $5 \mu\text{W}$  – estimated as 50X (same as size ratio) with the inverter pair dynamic power) [39].

TABLE IV. CIRCUIT COST ESTIMATION

Category	Synapse Order	0 <sup>th</sup> order	Higher order
	Metric		
Performance	Accuracy (%)	98	99
	Error (%)	2	1
Analog Implementation	Timescale Realization	1C (non-DPI)	1 or more C (non-DPI) or 1C and 3 biases (DPI)
	Driver Circuit	Series inverter pair	Op-Amp
	Voltage Margin Required	No	Yes
Digital Implementation	Quantization	Binary levels	4-bit 16 levels LUT
	Pulse generation	Counter	DAC

TABLE V. CRITICAL COMPONENTS: POWER AND AREA COMPARISON

Component	Higher Order	0 <sup>th</sup> order	Benefit
Driver [36][38]			
	Analog Buffer - OpAmp	Series Inverter Pair	
	Power: 100-500 $\mu\text{W}$ (Static)	Power: 100 nW (Dynamic)	1000 X
Pulse Generation [39][38]			
	DAC	Counter	
	Power: 500 $\mu\text{W}$	Power: 5 $\mu\text{W}$	100 X
	Area: 1500 $\mu\text{m}^2$	Area: 500 $\mu\text{m}^2$	3 X

All values are estimated for 0.25  $\mu\text{m}$  CMOS technology node 3.3V supply. Inverters and Analog buffers are assumed to drive 4pF load capacitance. All dynamic signals are assumed to be in ms (or 1kHz) range. 4bit Counters and 4bit DACs are assumed.

## V. CONCLUSION

In this paper, we performed a comprehensive parameter space exploration to show the correlation of the peak performance with an optimal reservoir spiking activity range for different synaptic orders and timescales in LSMs for the speech classification task. We demonstrated the true impact on the performance of the  $\delta$  synapse (96 %) compared to 2<sup>nd</sup> order synapse (99 %). We proposed the utility of 0<sup>th</sup> order synapses that are algorithmically a feasible choice (98 %) and practically an excellent circuit choice with numerous implementation related benefits. The elimination of large Op-Amps and power hungry digital-to-analog converter (DAC) circuits result in 2-3 orders of power and area savings when using 0<sup>th</sup> order synapses. Algorithmic simplifications to biologically plausible networks reduce the circuit burden while retaining the high performance of the classification task.

## ACKNOWLEDGMENT

VS thanks Maryam Shojaei Baghini and Ajay Singh for insightful discussions on circuit choices.

## REFERENCES

- [1] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Netw.*, vol. 10, no. 9, pp. 1659–1671, Dec. 1997, doi: 10.1016/S0893-6080(97)00011-7.
- [2] A. Taherkhani, A. Belatreche, Y. Li, G. Cosma, L. P. Maguire, and T. M. McGinnity, "A review of learning in biologically plausible spiking neural networks," *Neural Netw.*, vol. 122, pp. 253–272, Feb. 2020, doi: 10.1016/j.neunet.2019.09.036.
- [3] E. Painkras *et al.*, "SpiNNaker: A 1-W 18-Core System-on-Chip for Massively-Parallel Neural Network Simulation," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, Aug. 2013, doi: 10.1109/JSSC.2013.2259038.
- [4] M. Davies *et al.*, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan. 2018, doi: 10.1109/MM.2018.112130359.
- [5] A. Tavanaei and A. S. Maida, "Multi-layer unsupervised learning in a spiking convolutional neural network," *Proc. Int. Jt. Conf. Neural Netw.*, vol. 2017-May, pp. 2023–2030, 2017, doi: 10.1109/IJCNN.2017.7966099.
- [6] J. H. Lee, T. Delbruck, and M. Pfeiffer, "Training Deep Spiking Neural Networks Using Backpropagation," *Front. Neurosci.*, vol. 10, Nov. 2016, doi: 10.3389/fnins.2016.00508.
- [7] B. Rajendran *et al.*, "Specifications of nanoscale devices and circuits for neuromorphic computational systems," *IEEE Trans. Electron Devices*, vol. 60, no. 1, pp. 246–253, 2013, doi: 10.1109/TED.2012.2227969.
- [8] W. Maass, "Liquid State Machines: Motivation, Theory, and Applications," in *Computability in Context*, IMPERIAL COLLEGE PRESS, 2011, pp. 275–296, doi: 10.1142/9781848162778\_0008.
- [9] G. Srinivasan, P. Panda, and K. Roy, "SpiLinC: Spiking Liquid-Ensemble Computing for Unsupervised Speech and Image Recognition," *Front. Neurosci.*, vol. 12, p. 524, Aug. 2018, doi: 10.3389/fnins.2018.00524.
- [10] N. Soares and D. Kudithipudi, "Deep liquid state machines with neural plasticity for video activity recognition," *Front. Neurosci.*, vol. 13, no. JUL, pp. 1–12, 2019, doi: 10.3389/fnins.2019.00686.
- [11] A. Gorad, V. Saraswat, and U. Ganguly, "Predicting Performance using Approximate State Space Model for Liquid State Machines," in *2019 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2019, vol. 2019-July, pp. 1–8, doi: 10.1109/IJCNN.2019.8852038.
- [12] E. Goodman and D. Ventura, "Spatiotemporal Pattern Recognition via Liquid State Machines," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, Vancouver, BC, Canada, 2006, pp. 3848–3853, doi: 10.1109/IJCNN.2006.246880.
- [13] H. Ju, J.-X. Xu, E. Chong, and A. M. J. VanDongen, "Effects of synaptic connectivity on liquid state machine performance," *Neural Netw.*, vol. 38, pp. 39–51, Feb. 2013, doi: 10.1016/j.neunet.2012.11.003.
- [14] I. Chakraborty, G. Saha, and K. Roy, "Photonic In-Memory Computing Primitive for Spiking Neural Networks Using Phase-Change Materials," *Phys. Rev. Appl.*, vol. 11, no. 1, p. 1, 2019, doi: 10.1103/PhysRevApplied.11.014063.
- [15] S. Woźniak, A. Pantazi, T. Bohnstingl, and E. Eleftheriou, "Deep learning incorporating biologically inspired neural dynamics and in-memory computing," *Nat. Mach. Intell.*, vol. 2, no. 6, pp. 325–336, Jun. 2020, doi: 10.1038/s42256-020-0187-0.
- [16] S. R. Nandakumar, I. Boybat, M. Le Gallo, E. Eleftheriou, A. Sebastian, and B. Rajendran, "Experimental Demonstration of Supervised Learning in Spiking Neural Networks with Phase-Change Memory Synapses," *Sci. Rep.*, vol. 10, no. 1, p. 8080, Dec. 2020, doi: 10.1038/s41598-020-64878-5.
- [17] Y. Zhang, P. Li, Y. Jin, and Y. Choe, "A Digital Liquid State Machine With Biologically Inspired Learning and Its Application to Speech Recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2635–2649, Nov. 2015, doi: 10.1109/TNNLS.2015.2388544.
- [18] R. Legenstein and W. Maass, "6 What Makes a Dynamical System Computationally Powerful?," *New Dir. Stat. Signal Process.*, 2019, doi: 10.7551/mitpress/4977.003.0008.
- [19] C. Bartolozzi and G. Indiveri, "Synaptic dynamics in analog VLSI," *Neural Comput.*, vol. 19, no. 10, pp. 2581–2603, 2007, doi: 10.1162/neco.2007.19.10.2581.
- [20] W. Maass, T. Natschlager, and H. Markram, "Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations," *Neural Comput.*, vol. 14, no. 11, pp. 2531–2560, Nov. 2002, doi: 10.1162/089976602760407955.
- [21] A. Destexhe, Z. F. Mainen, and T. J. Sejnowski, "Kinetic models of synaptic transmission," *Methods Neuronal Model.*, vol. 2, pp. 1–25, 1998.
- [22] A. S. Lele, A. Naik, L. Bandhu, B. Das, and U. Ganguly, "Circuit Cost Reduction for Online STDP using NIPIN Selector as Timekeeping Device in RRAM Synapse," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, Sevilla, Oct. 2020, pp. 1–5, doi: 10.1109/ISCAS45731.2020.9180803.
- [23] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France, 1982, vol. 7, pp. 1282–1285, doi: 10.1109/ICASSP.1982.1171644.
- [24] D. Verstraeten, B. Schrauwen, D. Stroobandt, and J. Van Campenhout, "Isolated word recognition with the Liquid State Machine: a case study," *Inf. Process. Lett.*, vol. 95, no. 6, pp. 521–528, Sep. 2005, doi: 10.1016/j.ipl.2005.05.019.
- [25] T. Chavan, S. Dutta, N. R. Mohapatra, and U. Ganguly, "Band-to-Band Tunneling Based Ultra-Energy-Efficient Silicon Neuron," *IEEE Trans. Electron Devices*, vol. 67, no. 6, pp. 2614–2620, 2020, doi: 10.1109/TED.2020.2985167.
- [26] S. Dutta, V. Kumar, A. Shukla, N. R. Mohapatra, and U. Ganguly, "Leaky Integrate and Fire Neuron by Charge-Discharge Dynamics in Floating-Body MOSFET," *Sci. Rep.*, vol. 7, no. 1, pp. 1–7, 2017, doi: 10.1038/s41598-017-07418-y.
- [27] B. Das, A. Lele, P. Kumbhare, J. Schulze, and U. Ganguly, "Pr x Ca 1–x MnO 3 -Based Memory and Si Time-Keeping Selector for Area and Energy Efficient Synapse," *IEEE Electron Device Lett.*, vol. 40, no. 6, pp. 850–853, Jun. 2019, doi: 10.1109/LED.2019.2914406.
- [28] A. Fumarola *et al.*, "Bidirectional Non-Filamentary RRAM as an Analog Neuromorphic Synapse, Part II: Impact of Al/Mo/Pr0.7Ca0.3MnO3 Device Characteristics on Neural Network Training Accuracy," *IEEE J. Electron Devices Soc.*, vol. 6, no. 1, pp. 169–178, 2018, doi: 10.1109/JEDS.2017.2782184.
- [29] I. Sourikopoulos *et al.*, "A 4-TJ/Spoke Artificial Neuron in 65 nm CMOS Technology," *Front. Neurosci.*, vol. 11, Mar. 2017, doi: 10.3389/fnins.2017.00123.
- [30] S. Lashkare, S. Chouhan, T. Chavan, A. Bhat, P. Kumbhare, and U. Ganguly, "PCMO RRAM for Integrate-and-Fire Neuron in Spiking Neural Networks," *IEEE Electron Device Lett.*, vol. 39, no. 4, pp. 484–487, Apr. 2018, doi: 10.1109/LED.2018.2805822.
- [31] S. Dutta, T. Bhattacharya, N. R. Mohapatra, M. Suri, and U. Ganguly, "Transient variability in SOI-Based LIF neuron and impact on



- unsupervised learning,” *IEEE Trans. Electron Devices*, vol. 65, no. 11, pp. 5137–5144, Nov. 2018, doi: 10.1109/TED.2018.2872407.
- [32] G. M. Bo, D. D. Caviglia, and M. Valle, “An on-chip learning neural network,” in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, 2000, vol. 4, pp. 66–71 vol.4, doi: 10.1109/IJCNN.2000.860751.
- [33] A. Shukla, V. Kumar, and U. Ganguly, “A software-equivalent SNN hardware using RRAM-array for asynchronous real-time learning,” in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, vol. 2017-May, pp. 4657–4664, doi: 10.1109/IJCNN.2017.7966447.
- [34] M. D. N. Mehta and M. P. N. Sanghavi, “Intel’s 45nm CMOS Technology Performance Parameters in VLSI Design,” vol. 4, p. 3, 2013.
- [35] “Intel’s 45nm CMOS Technology,” *Intel Technol. J.*, vol. 12, no. 2, p. 90, 2008. doi: 10.1535/itj.1202
- [36] R. Z. Shi and T. Horiuchi, “A Summating, Exponentially-Decaying CMOS Synapse for Spiking Neural Systems,” in *Advances in Neural Information Processing Systems*, p. 1003-1010, 2003
- [37] F. Maloberti, Ed., “CMOS Operational Amplifiers,” in *Analog Design for CMOS VLSI Systems*, Boston, MA: Springer US, 2001, pp. 217–324. doi: 10.1007/0-306-47952-4\_5
- [38] J. Mahattanakul and J. Chutichatuporn, “Design procedure for two-stage CMOS opamp with flexible noise-power balancing scheme,” *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 52, no. 8, pp. 1508–1514, Aug. 2005, doi: 10.1109/TCSI.2005.851395.
- [39] S. L. Harris and D. M. Harris, “5 - Digital Building Blocks,” in *Digital Design and Computer Architecture*, S. L. Harris and D. M. Harris, Eds. Boston: Morgan Kaufmann, 2016, pp. 238–293. doi: 10.1016/B978-0-12-800056-4.00005-4
- [40] T. Y (Yongjian), “Smart and high-performance digital-to-analog converters with dynamic-mismatch mapping,” 2010, doi: 10.6100/IR685413.
- [41] D. Khilwani *et al.*, “PrxCa1-xMnO3 based stochastic neuron for Boltzmann machine to solve ‘maximum cut’ problem,” *APL Mater.*, vol. 7, no. 9, p. 091112, Sep. 2019, doi: 10.1063/1.5108694.
- [42] K. Chander and S. Choudhry, “65nm Low Power Digital to Analog Converter for CUWB,” in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, May 2018, pp. 610–614, doi: 10.1109/ICOEI.2018.8553815.