



OPEN Hopf physical reservoir computer for reconfigurable sound recognition

Md Raf E. Ul Shougat^{1,5}, XiaoFu Li^{2,5}, Siyao Shao^{3,4,5}, Kathleen McGarvey^{3,5} & Edmon Perkins^{2,5}✉

The Hopf oscillator is a nonlinear oscillator that exhibits limit cycle motion. This reservoir computer utilizes the vibratory nature of the oscillator, which makes it an ideal candidate for reconfigurable sound recognition tasks. In this paper, the capabilities of the Hopf reservoir computer performing sound recognition are systematically demonstrated. This work shows that the Hopf reservoir computer can offer superior sound recognition accuracy compared to legacy approaches (e.g., a Mel spectrum + machine learning approach). More importantly, the Hopf reservoir computer operating as a sound recognition system does not require audio preprocessing and has a very simple setup while still offering a high degree of reconfigurability. These features pave the way of applying physical reservoir computing for sound recognition in low power edge devices.

There are ubiquitous methods of audio signal classification, particularly for speech recognition^{1,2}. However, machine learning suffers several drawbacks that hinder its wide dissemination on the Internet of Things (IoT)³. First, machine learning, especially deep neural networks (DNNs), rely on the cloud infrastructure to conduct massive computation for both model training and inference. State-of-the-art (SOTA) deep learning models, such as GPT-3, can have over 175 billion parameters and training requirements of 3.14×10^{23} FLOPS (floating operations per second)^{4,5}. The training of the SOTA speech transcription model, Whisper, used a word library that had as many words as one person would continuously speak for 77 years⁶. None of these mentioned technical requirements could be fulfilled by any edge devices of IoT; thus, the cloud infrastructure is a necessity for DNN tasks. Second, reliance of cloud computing for machine learning poses great security and privacy risks. Over 60% of previous security breaches happened during the raw data communication between the cloud and the edge for machine learning⁷. Further, each breach carries an average \$4.24 million loss, and this number is continuously growing⁸. The privacy concern causes distrust among smart device users and drives the abandonment of smart devices^{9,10}. Third, the environmental impact of implementing DNN through a cloud infrastructure is often overlooked but cannot be neglected. Training a transformer model with 213 million parameters will generate carbon dioxide emissions equaling four times of a US manufacturer's vehicle over its whole lifespan¹¹. Therefore, the next generation of smart IoT devices needs to possess sufficient computational power to operate machine learning or even deep learning on the edge.

Among efforts to bring machine learning to edge devices, reservoir computing, especially physical reservoir computing, has generated early success over the last two decades. Originating from the concepts of liquid state machines and echo state networks, researchers demonstrated that the sound-induced ripples on the surface of a bucket of water could be used to conduct audio signal recognition¹². In a nutshell, reservoir computing exploits the intrinsic nonlinearity of a physical system to replicate the process of nodal connections in a neural network to extract features from time series signals for machine perception^{13,14}. Reservoir computing directly conducts computations in an analog fashion by using the physical system, which largely eliminates the necessity of separate data storage, organization, and machine learning perception. Notably, reservoir computing is naturally suited for audio processing tasks, which are a subset of time series signals.

Researchers have explored many physical systems to operate as reservoir computers for temporal signal processing. These systems include the field-programmable gate array (FPGA)¹⁵, chemical reactions¹⁶, memristors¹⁷,

¹Mechanical & Aerospace Engineering Department, North Carolina State University, 1840 Entrepreneur Drive, Raleigh, NC 27695, USA. ²LAB2701, Atwood, OK 74827, USA. ³TandemLaunch, 780 Av. Brewster, Montreal H4C2K1, Canada. ⁴echosonic, 780 Av. Brewster, Montreal H4C2K1, Canada. ⁵These authors contributed equally: Md Raf E. Ul Shougat, XiaoFu Li, Siyao Shao, Kathleen McGarvey and Edmon Perkins. ✉email: edmon@lab2701.com

superparamagnetic tunnel junctions¹⁸, spintronics¹⁹, attenuation of wavelength of lasers in special mediums²⁰, MEMS (microelectromechanical systems)²¹, and others^{13,22}. Though these studies have demonstrated that reservoir computing could handle audio signal processing, the physical system for computing is usually very cumbersome²⁰, and they all require preprocessing of the original audio clips using methods such as the Mel spectrum, which largely cancels the benefits of reducing the computational requirements of machine learning via reservoir computing. More importantly, to boost the computational power, conventional reservoir computing techniques use time-delayed feedback achieved by a digital to analog conversion²³, and the time-delayed feedback will hamper the processing speed of reservoir computing while drastically increasing the envelope of energy consumption for computing. We suggest that the less-than-satisfactory performance of physical reservoir computing is largely caused by the insufficient computational power of the computing systems chosen by the previous works.

Recently, we have discovered that the Hopf oscillator, which is a common model for many physical processes, has sufficient computational power to conduct machine learning. Although this is a very simple physical system, computing can be achieved without the need of additional data handling, time-delayed feedback, or auxiliary electrical components^{24–27}. Interestingly, the nonlinear activation of a neural network can also sometimes be captured by the physical reservoir, which can further simplify the physical reservoir computer's architecture (e.g., a shape memory alloy actuator physical reservoir computer²⁸). The performance of the Hopf oscillator reservoir computer on a set of benchmarking tasks (e.g., logical tasks, emulation of time series signals, and prediction tasks) is exceptional compared to much more complex physical reservoirs.

This paper is an extension of previous work to further demonstrate the outstanding capabilities of the Hopf reservoir computer for audio signal recognition tasks. The Hopf oscillator acts as a nonlinear filter, but also some portion of the computational task is off-loaded to the Hopf physical reservoir computer. Based on our previous work, the Hopf oscillator both performs computation and stores information in its dynamic states^{24,25}. Fundamentally, the nonlinear response of the oscillator is a type of nontraditional computing, which is unlocked through machine learning. Further, the oscillator's dynamic states act as a type of local memory, as no additional memory was introduced through delay lines. In this previous work on the Hopf oscillator, a single readout layer was trained to perform a battery of tasks. Here, the single readout layer is replaced with a relatively shallow neural network for more difficult tasks, such as sound recognition. These results point to the efficacy of using this type of reservoir computer for edge computing, which could pave the way to obtaining edge artificial intelligence and decentralized deep learning in the foreseeable future.

Hopf oscillator and reservoir

The forced Hopf oscillator is represented in Eq. (1)^{27,29}:

$$\begin{aligned}\dot{x} &= (\mu - (x^2 + y^2))x - \omega_0 y + A \sin(\Omega t) \\ \dot{y} &= (\mu - (x^2 + y^2))y + \omega_0 x\end{aligned}\quad (1)$$

In the above equations, x and y refer to the first and second states of the Hopf oscillator, respectively. The ω_0 term is the resonance frequency of the Hopf oscillator. The μ parameter affects the radius of the limit cycle motion. For example, without external forcing, the Hopf oscillator would have a limit cycle of radius μ , and it would oscillate at a frequency of ω_0 . This parameter also loosely correlates with the quality factor of the oscillator. A is the amplitude of a sinusoidal force.

For the oscillator to classify audio signals, an external forcing signal that contains the audio signal, $a(t)$ is constructed, which is shown in Eq. (2); this is then used as input to the Hopf oscillator. The modified Hopf oscillator as a reservoir is represented by Eqs. (3) and (4):

$$f(t) = 1 + a(t) \quad (2)$$

$$\dot{x} = (\mu f(t) - (x^2 + y^2))x - \omega_0 y + A f(t) \sin(\Omega t) \quad (3)$$

$$\dot{y} = (\mu f(t) - (x^2 + y^2))y + \omega_0 x \quad (4)$$

The external signal, $f(t)$, is composed of a DC offset and the audio signal, $a(t)$. The DC offset ensures that the radius parameter is non-negative. This external signal is injected into both the radius parameter, μ , and the sinusoid, $A \sin(\Omega t)$. The Hopf oscillator dynamically responds to the audio signal, and the x state corresponds to the audio features for the machine learning audio classification task. The y state, although not explicitly used in the classification task (as depicted in Fig. 1), likely stores information and aids in the computational task. Unlike the original form of the Hopf oscillator reservoir computer, we use the Hopf oscillations to extract audio features for classification instead of directly using the two state outputs for time series signal prediction²⁴. As such, several changes are made in the computational scheme of the Hopf oscillator reservoir computer. First, this formulation of the reservoir does not include the typical procedure of multiplication of inputs with the masking function, as no masking function is included. Conventional reservoir computing uses a preset mask multiplying the reservoir outputs to create neurons in the reservoir system. The training of the mask equates to updating parameters when training the digitally realized neural networks. However, this method is memory expensive and inefficient for audio signal processing, since the length of the mask should be sufficient to cover the length of audio clip and the nodal connections necessary for the signal classification. Instead of training masks, we use a more efficient multiple layer convolutional neural network readout to directly feed forward the reservoir outputs and train the connections between each layer as the parameters. Second, Gaussian noise is not multiplied with

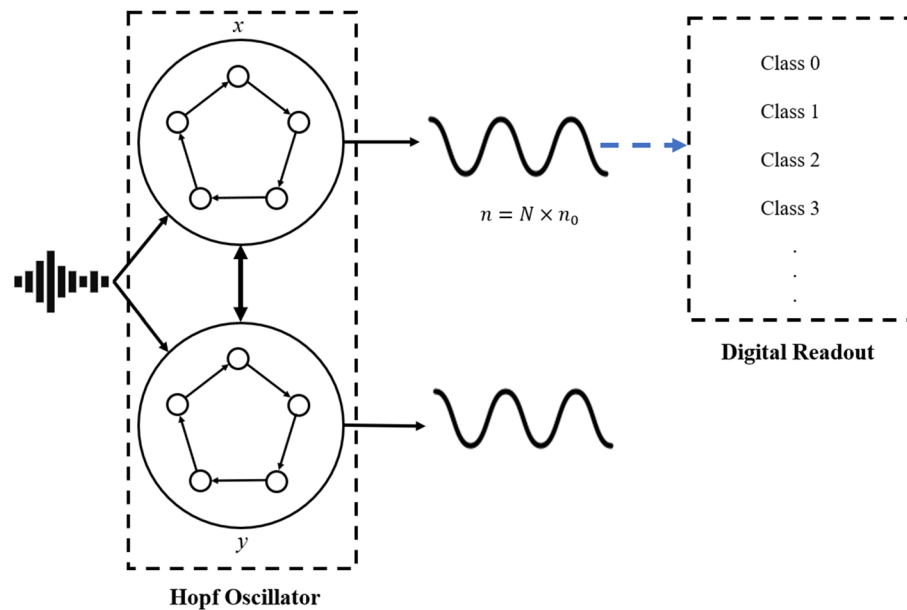


Figure 1. A schematic showing the nodal connections within a Hopf oscillator for reservoir computing. The original signal, $f(t)$, is sent to the two states of the oscillator (i.e., two physical nodes). Each physical node generates N virtual nodes in time series. The digital readout layers (i.e., machine learning algorithm) will read n samples from the node x of the oscillator (note that we only use one node for audio classification in the present paper). n_0 corresponds to the number of samples from the original audio signal, and N refers to the number of virtual nodes controlled by the readout mechanisms. The signal from the reservoir is then sent to a neural network, which is indicated by the blue dashed arrow; this neural network is described in Fig. 12. The digital readout will classify the n samples corresponding to one audio clip to its class.

the audio signal, as the audio signals already have background noise. This noise mask was used in a previous Hopf reservoir computer study to highlight its robustness²⁴. Third, instead of using a pseudo-period to guide the training of the machine learning readout, we use the number of samples collected for classification to control the nodal connections within each collected feature point generated from the reservoir processing 1D audio data. N virtual nodes mean that for each sampling point of the original audio, the reservoir will generate $N - 1$ nodal connections in 1D for each reservoir state for classification. For example, with N virtual nodes, a sampled audio data point is processed by the physical node (i.e., x in Fig. 1) $N - 1$ times, which creates N feature points from one audio sample and $N - 1$ nodal connections in these N feature points. In the current paper, we set N to 100 for audio processing. This method hinders the sampling speed of the audio signals. Thus, we resample the original full resolution audio data to ensure that we operate experiments within a relatively short period of time. It is worth noting that the length of the audio clips for each classification event effectively builds the pseudo-period in the traditional context of the reservoir computing via time-delayed feedback loops (i.e., a fixed length of the audio will produce one classification result with details provided later). The eventual nodal connection of the Hopf reservoir computer and output handling could be conceptualized as Fig. 1.

Here, the Hopf reservoir computer is used to compute feature maps, with several representative examples shown in Fig. 2. “VN#” refers to the virtual node number, and the time scale for the other axis is defined such that the step size is the reciprocal of the sampling rate. The value of the feature map is rescaled from 0 to 1. Consecutive convolutional layers, followed by the flattened layer and fully-connected layers depicted in Fig. 12, construct the machine learning readout for processing the audio signal outputs from the reservoir, which is further described in “Methods” Section. Note that a similar approach is applied in the SOTA urban sound recognition on edge devices³⁰, though we eliminate the computationally expensive preprocessing of the Mel spectrogram by offloading feature extraction to the reservoir computer. More importantly, our approach could use a very coarse sampling (4000 Hz was used here) instead of the Mel spectrogram applied in³⁰ to capture the granularity of the audio signals. A detailed comparison is provided in the subsequent section to demonstrate the superior feature extraction from the Hopf reservoir computer.

Results

Results for urban sound recognition dataset. First, we present the results of the Hopf reservoir computer for an urban sound recognition task. As shown in Fig. 3 in the left column, the audio features from the Mel spectrum operations (as calculated on the audio clips with a 44.1 kHz sampling rate) show drastic differences between the three examples; using the top example as a reference, the average pointwise Euclidean distance between the reference and the other two are higher than 25. In comparison, the audio features from the Hopf RC

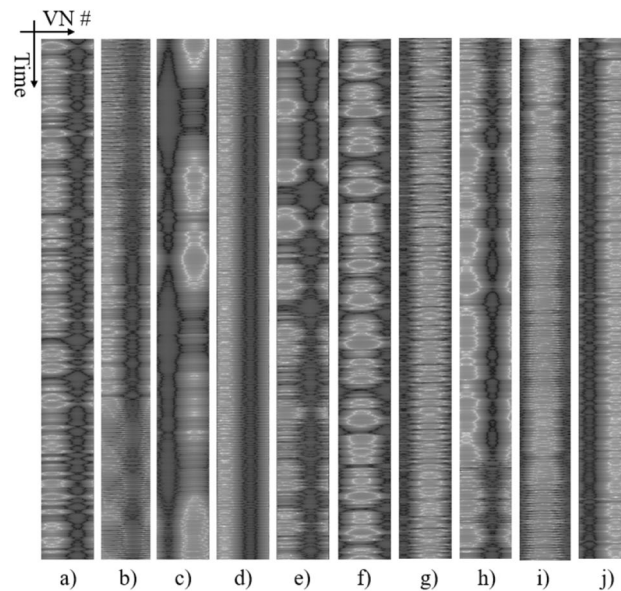


Figure 2. Sample feature maps generated by the Hopf oscillator corresponding to different audio events. Each audio clip has a length of 1 sec sampled at 4000 Hz. The x-axis follows the arithmetic order of the virtual nodes, and the y-axis is the time. The reservoir is set to have 100 nodes for the test. The grayscale value (from 0 to 1) of each pixel corresponds to the signal strength of each data point (i.e., feature point of the audio signal). (a) Air conditioner. (b) Car horn. (c) Children playing. (d) Dog barking. (e) Drilling. (f) Engine idling. (g) Gunshot. (h) Jackhammer. (i) Siren. (j) Street music.

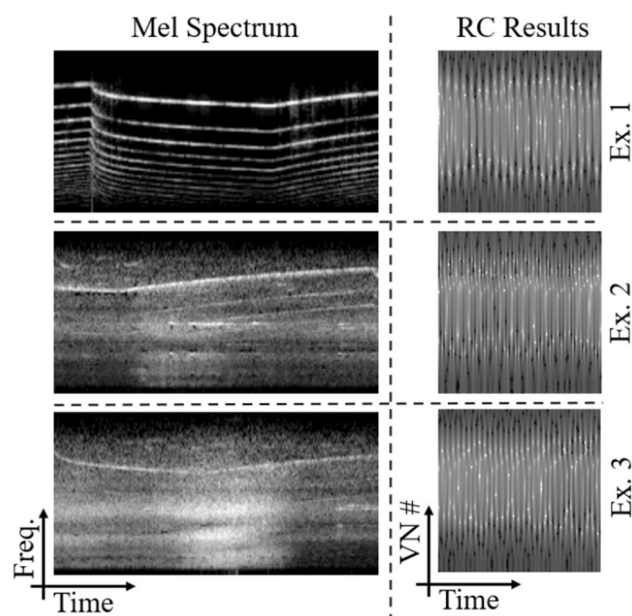


Figure 3. The Mel spectrum is compared with the Hopf RC for the urban sound recognition task. From the top to the bottom, three examples of the siren class are presented. In the left column, the energy of the Mel spectrum is shown, where the horizontal axis is time and the vertical axis is frequency. The Mel spectrum operation is conducted upon samples that are four seconds long with a 44.1 kHz sampling rate. The total number of frequency bands is set to be 100, and the time step is set to be 0.025 seconds. In the right column, the audio features extracted from the Hopf reservoir computer for the same samples, such that each 1 second audio clip is downsampled to 4000 Hz and the number of virtual nodes is set to 100. Notably, the Mel results and the Hopf reservoir results do not look similar to each other, but the information conveyed by each process is internally consistent, which is highlighted by the classifier's performance.

are shown in the right column of Fig. 3; all three examples have a much higher similarity for these three examples (e.g., Euclidean distance < 12). The average Euclidean distance for the samples between classes is:

$$\frac{\sum_{i,j} \left(\frac{\sum_{\alpha,\beta} \sqrt{(c_i(x_\alpha, y_\beta) - c_j(x_\alpha, y_\beta))^2}}{(N(\text{number of seconds})(\text{sampling rate}))} \right)}{(\text{samples in class } I)(\text{samples in class } J)} \quad (5)$$

where $c(x, y)$ is the amplitude of the Hopf reservoir computer at time x and virtual node number y . Here, i is indexed over class I , j is indexed over class J , α is indexed over all the values of x , and β is indexed over all the values of y . The average Euclidean distances are presented in Fig. 4. The diagonal has the minimal value for each column and row, which demonstrates that the Hopf oscillator is capable of separating the classes, even without the neural network.

a)	10.6									
b)	22.3	14.1								
c)	20.5	28.4	10.8							
d)	33.8	19.7	26.3	10.4						
e)	21.6	33.5	24.8	27.3	12.9					
f)	27.4	25.0	31.9	25.9	24.5	11.6				
g)	32.6	36.7	31.6	27.7	27.5	24.4	13.7			
h)	26.5	27.1	20.7	23.9	30.3	19.7	21.5	12.3		
i)	32.7	28.4	18.3	20.8	25.1	29.1	30.0	24.9	9.7	
j)	29.3	23.5	24.1	25.9	30.4	29.3	27.7	19.7	23.7	11.5
	a)	b)	c)	d)	e)	f)	g)	h)	i)	j)

Figure 4. The average Euclidean distances are presented in this symmetric matrix for the ten urban sound classes. The average Euclidean distances were calculated between all the samples within a single class (diagonal elements) and between all the samples from two classes (off-diagonal elements).

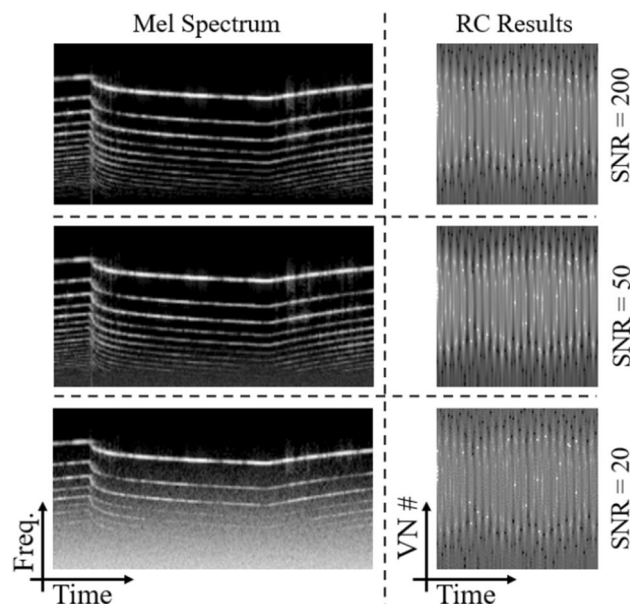


Figure 5. The robustness of the Hopf RC audio extraction is compared with the Mel spectrum for various signal-to-noise ratios (SNRs). For visualization, the siren example shown in the top of Fig. 3 is used with different levels of noise. From the top to the bottom, three different amounts of noise were added to the original siren audio example. In the left column, the energy of the Mel spectrum is shown. Note that the result starts to lose low frequency information when the SNR drops to 20. In the right column, the audio features that are extracted using the Hopf RC are shown. Note that the result remains largely the same for all noise levels, even when the SNR is equal to 20.

The robustness of the audio classification is also of high importance for real-world applications. To highlight this, the Mel spectrum results are compared with the Hopf RC results for three different noise levels. Using the example in the top row of Fig. 3, white noise is added to the original signal to create different signal-to-noise ratios (SNRs); the audio features of these three new signals are computed with the Mel spectrum (using 44.1 kHz audio sampling rate) and the Hopf reservoir computer (using 4000 Hz audio sampling rate). The output audio features are shown in Fig. 5. It is clearly shown that the Mel spectrum-based audio features lose low frequency information when the SNR is reduced to 20, while the features generated by the Hopf reservoir computer maintain a similar structure with the original audio counterpart, with the Euclidean distance < 5 for an SNR of 20.

The confusion matrix for the urban sound recognition task is shown in Fig. 6. The proposed audio recognition approach based on the Hopf reservoir computer has a 96.2% accuracy. This accounts for a 10% accuracy improvement compared to³⁰, with a reduction of $> 94\%$ of the FLOPS (floating operations per second) for high sampling rate readout and Mel spectrum computation and $\sim 90\%$ of the audio pieces for training.

Results for Qualcomm voice command dataset. Using the machine learning model trained from the previous test case (i.e., the urban sound recognition task) as the baseline, we test the Qualcomm voice command dataset to demonstrate the reconfigurability of the Hopf reservoir computer audio recognition system. In this experiment, we purposefully reduce the number of epochs to 20 and freeze the CNN portion of the machine learning model to reconfigure the process of the audio recognition system from the urban sound detection task to a voice command task. In the left portion of Fig. 7, representative audio features of the four classes are shown, which have significant differences compared to the features of the urban sound events (Fig. 2). The audio recognition yields a $> 99\%$ accuracy, with the confusion matrix depicted in the right portion of Fig. 7. Note that the number of parameters trained for this experiment is about 35,000, which accounts for about 300 KB dynamic memory for 8-bit input with a batch size of 5^{31,32}, demonstrating the feasibility of running the training of the machine learning readout on low-level edge devices consuming Li-Po battery level of power.

Results for spoken digit dataset. The spoken digit dataset is used to compare the performance of the Hopf reservoir computer for audio recognition with other reservoirs (e.g.,^{15–22}). As shown in Fig. 8, the Hopf reservoir computer produces an approximately 97% accuracy for the spoken digit classification task. This result retains the state-of-the-art recognition accuracy on this dataset while only using one physical device (i.e., one consolidated analog circuit) and two physical nodes (x and y states). As a comparison, the best performing reservoir¹⁷ employed 10 memristors and preprocessing of the original audio clips to yield a similar accuracy. We suggest that the vibratory nature of our reservoir largely contributes to the simplicity of the proposed sound event detection system, and the activation of the reservoir using sinusoidal signals boosts the feature extraction of the audio signal using Hopf oscillations (details described later).

Further, we increase the strength of the activation signal (term A in Eq. 1) and discard the inverse hyperbolic tangent activation (Eq. 6) before the machine learning readout. The yielded results, which are shown in Fig. 9, have a 96% accuracy compared to the case using Eq. (6) before sending the x state to the machine learning readout. This suggests that this Hopf reservoir computer can be reconfigured by its digital readout, similar to other physical reservoir computers. Additionally, the Hopf oscillator's computational power could also be drastically enhanced by changing the oscillator's internal physical conditions.

True class	a)	95.8%	0.3%	0.5%	0.6%	0.4%	0.4%	0.8%	0.4%	0.1%	0.6%
	b)	0.2%	96.1%	0.4%	0.2%	0.6%	0.7%	0.7%	0.3%	0.4%	0.2%
	c)	0.1%	0.2%	96.9%	0.2%	0.5%	0.5%	0.4%	0.5%	0.1%	0.5%
	d)	0.6%	0.5%	0.6%	96.4%	0.2%	0.1%	0.5%	0.4%	0.1%	0.5%
	e)	0.2%	0.4%	0.4%	0.2%	96.5%	0.5%	0.8%	0.3%	0.2%	0.4%
	f)	0.1%	0.3%	0.5%	0.9%	0.1%	96.7%	0.2%	0.2%	0.3%	0.6%
	g)	0.5%	0.3%	0.3%	0.6%	0.2%	0.6%	96.2%	0.5%	0.3%	0.3%
	h)	0.5%	0.6%	0.6%	0.6%	0.4%	0.7%	0.4%	95.3%	0.6%	0.2%
	i)	0.7%	0.5%	0.5%	0.2%	0.4%	0.4%	0.6%	0.4%	96.1%	0.1%
	j)	0.2%	0.4%	0.6%	0.7%	0.5%	0.6%	0.4%	0.4%	0.2%	95.7%
		a)	b)	c)	d)	e)	f)	g)	h)	i)	j)
		Predicted class									

Figure 6. For the urban sound recognition task, the confusion matrix is presented with the recognition accuracy labeled for the ten different audio events. Note that the class labels in this figure are the same as the class labels of Fig. 2.

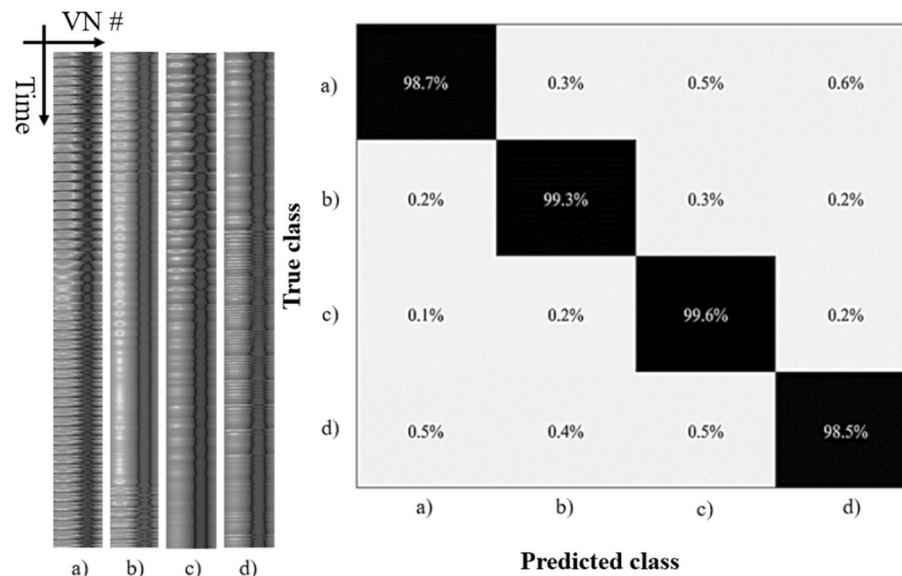


Figure 7. Summary of the results of the Hopf reservoir computer for the Qualcomm voice command task. Left: Examples of the feature maps of different wake words generated by the Hopf reservoir computer. Right: The confusion matrix of the proposed sound recognition system processing Qualcomm wake words. Each label corresponds to: (a) “Hi, Galaxy”, (b) “Hi, Lumia”, (c) “Hi, Snapdragon”, and (d) “Hi, Android”.

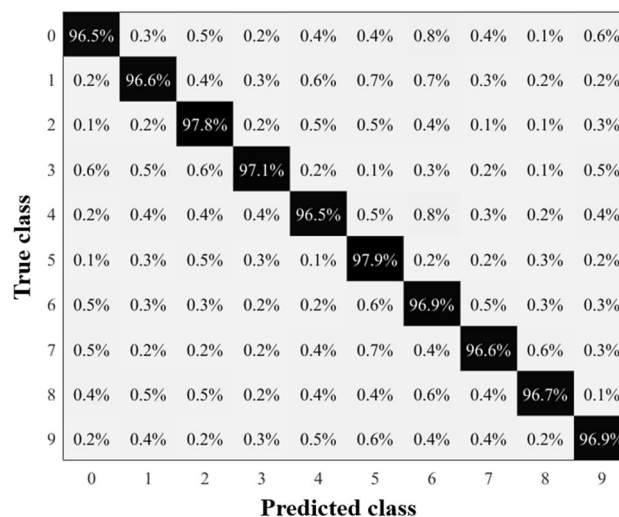


Figure 8. Summary of the results of Hopf reservoir computer conducts spoken digits recognition task. The confusion matrix of the proposed sound recognition system processing spoken digits dataset with original activation strength and inverse hyperbolic tangent before machine learning readouts.

Conclusions

This Hopf physical reservoir computer architecture is proposed for real-world edge computing applications, such as audio recognition. Although speech recognition is a relatively simple task for deep neural networks running on the cloud, it is a difficult task for edge computers due to their limited computational power. The proposed architecture effectively uses the strengths of both analog and digital devices by splicing an analog oscillator to a digital neural network. Moreover, the Hopf oscillator can be readily fabricated from commercial off the shelf electrical components.

The Hopf physical reservoir computer architecture discussed in this paper has several distinct differences from other similar physical reservoir computers. Most prominently, this Hopf oscillator is paired with a neural network rather than using a simple ridge regression. By increasing the complexity of the neural network, the Hopf physical reservoir computer is able to perform more difficult tasks. As the neural network is straightforward, it can be easily implemented. The architecture employed in this paper does not use any preprocessing of the original

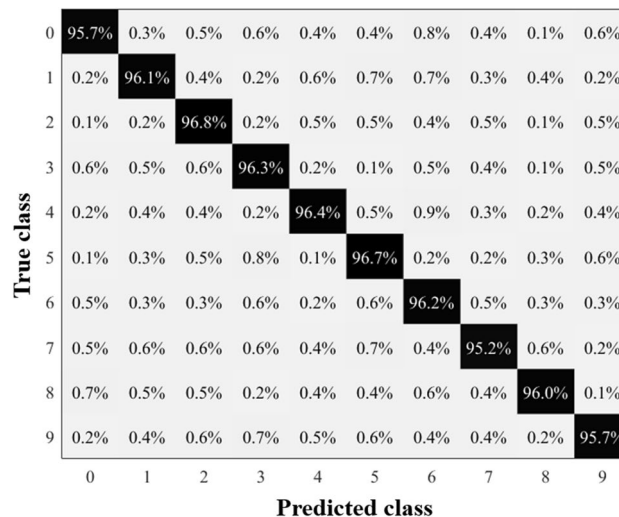


Figure 9. Summary of the results of Hopf reservoir computer conducts spoken digits recognition task. The confusion matrix of the proposed sound recognition system processing spoken digits dataset with 10 times increase of activation strength and without inverse hyperbolic tangent before machine learning readouts.

audio data, which significantly reduces the computational costs of the recognition task. Instead, it follows the activation signal to construct the feature maps by matrix reshape and inverse tanh. Usually, the Mel spectrum is used for this type of task, which can account for more than half of the computational load³³. Most nonlinear oscillator-based physical reservoir computers must use time-delayed feedback, which is cumbersome as it would require digital-to-analog and analog-to-digital converters. However, the Hopf oscillator is capable of storing enough information in its dynamic states to avoid this^{24,25}. Moreover, the presented architecture is robust to noise because of the Hopf oscillator's nonlinearity, which is important for real-world audio processing applications.

The proposed architecture has several key advantages. First, the computational load for the proposed approach is significantly reduced. The computations involved in the construction of the feature maps are matrix reshape, normalization, and inverse tanh. These operations only consume around 10% of the computational power compared to the Mel spectrogram for a sampling rate of 4,000 Hz. An estimate of the computational load draws the conclusion that similar operations on Cortex-M4 (Arm, San Jose, California) edge devices yields only about 5 ms of the latency running this algorithm. Second, the proposed method can be paired with different machine learning models. Though the paper uses the CNN as the machine learning readout, the feature map yielded from the proposed method can be replaced by common image processing methods, including but not limited to transformer⁽³⁴⁾, structure similarity index⁽³⁵⁾, feedforward neural network⁽³⁶⁾, and Euclidean distance⁽³⁷⁾, etc. Third, compared to the Mel spectrogram, physically implemented limit cycles can generate features that are robust to both noise and low audio quality. It is worth noting that the audio used for the experiments is a downsampled version, which is about half of the sampling rate used by the Mel + CNN approach, while still achieving an audio recognition accuracy that is approximately 10% higher. As an example of this robustness, the feature map generated from the audio with additional noise (Fig. 5) retains its distinctive features even under extremely low signal-to-noise ratio (< 20).

Summary of results. In this paper, we present the results of sound signal recognition using reservoir computing technology consisting of a Hopf oscillator^{24,25}. Instead of employing computationally expensive preprocessing (e.g., Mel spectrum) commonly used in other studies^{15,17,20,30}, we directly take the outputs from the Hopf circuit to process the normalized audio signal for machine learning recognition. We anticipate that this Hopf reservoir computing can be directly implemented to microphones to achieve a future processing-on-the-sensor.

In “Results” Section, we systematically demonstrate that our Hopf reservoir computing approach yields a 10% accuracy improvement on a diverse 10-class urban sound recognition compared to the state-of-the-art results using edge devices³⁰, whereas we use a surprisingly simple preprocessing by just normalizing the original signal. The wake words recognition results in > 99% accuracy using the exact readout machine learning algorithm by only retraining the MLP. This implies that the Hopf reservoir computer will enable inference and reconfiguration on the edge for the sound recognition system. Additionally, compared to other reservoir computing systems (e.g.,^{15–17,22}), the spoken digit dataset yields superior performance without the need of using complex preprocessing, multiple physical devices, or mask functions; in addition, we have also conducted our benchmarking experiments on far more realistic datasets (i.e., the 10-class urban sound recognition dataset and the 4-class wake words dataset). We demonstrate boosted performance of audio signal processing by changing the activation signal strength of the Hopf oscillator, which implies that there are more degrees of freedom for reconfiguring physical reservoir computers as compared to other reservoir implementations.

Lastly, we carefully crafted the algorithms and preprocessing of the data for sound recognition tasks to keep overall energy consumption, including the digital readout, less than 1 mW based on FLOPS operations and

the analog sampling rate. The computational load, which uses less than 700 sound clips of a 10-class dataset for training machine learning models, is well below the envelope of the computational resources possessed by consumer electronic devices. As such, the sound recognition devices using a Hopf reservoir computer could have an effortless integration with devices with untraceable computational load increases.

Analysis on the physical mechanisms of the Hopf reservoir computer sound recognition. Three elements play important roles in the audio signal recognition. The limit cycle system creates an oscillation signal in the temporal domain with a sinusoidal form, which continuously convolves with the incoming audio signal. This convolution is reminiscent of the Fourier transform, and the Hopf oscillator generates unique patterns for audio recognition (e.g., Fig. 2). Interestingly, this process largely replicates the process of the cochlea in extracting the sound signal features perceptible by the neurons. The nonlinear oscillation of the Hopf oscillator in the temporal direction creates nodal connections of the reservoir computer, corresponding to the neuron connections in DNN. Additionally, the nonlinearity of the Hopf oscillator causes it to respond differently to signals possessing various characteristic features of the audio in a broadband fashion, which produces clean separation of features (Figs. 2 and 7a). It is worth noting that some recent studies^{38,39} have demonstrated that the cochlea and its directly-connected neurons create a limit cycle system using the previous audio signals as activation to dynamically enhance the performance of the cochlea in performing audio signal feature extraction. The physical model of the inner ear can be modeled as a Hopf oscillator with a time-delayed feedback loop using the signals from previous time instants to activate the limit cycle oscillations. The audio signal recognition actually happens in the inner ear instead of in the brain. An interesting future extension of this work is to explore different activation signals to create an artificial ear, which is capable of on-membrane audio recognition. In the meantime, the two states of the Hopf oscillator affect each other with a time delay, which enhances the memory effects essential to the time series signal processing.

Discussion and future work. The unique advantages of the Hopf reservoir computer demonstrated in this paper pave the way for the next generation of smart IoT devices that exploit the unused computational power in sensor networks. Specifically, the physical mechanisms backing reservoir computing also happen in the microphone membrane with carefully crafted activation signals³⁸. One could imagine that future microphones directly operate sound signal recognition using sensor mechanisms instead of dedicated processing rigs. In addition, as shown in Fig. 2, the feature map of sound signals consists of unique patterns that are recognized by a convolutional neural network commonly used for visual signal processing. An extension of the present work will explore the correlations of audio signal feature maps, visual signal feature maps, and other types of time series data features. As such, reservoir computing could be used as a backbone for multi-modal machine learning in smart IoT paradigms, including sensor fusion, audio video signal combination, and decentralized machine learning. The extremely small amount of training data required for the machine learning operation and clear feature separation described in “Results” Section could offer surprisingly satisfactory results, which is essential for many use cases without the luxury of unlimited sizes of datasets (e.g., soft user identification) or with noisy environments (e.g., a mix of different signals). One example is shown in Fig. 10: a eight-second long audio signal consisting of multiple different (i.e., car horn, drilling, and siren) is used to demonstrate the proof-of-concept of Hopf reservoir computer on mixed signal processing. The first four seconds of the audio clip only have car horn and drilling sound. For the last four seconds, the siren sound is added with a higher amplitude. As shown in the figure, the audio features generated from the Hopf reservoir computer has a clearly dominant class on the second half of the data and exhibits visually high correlation with the audio features generated by a clean siren sound

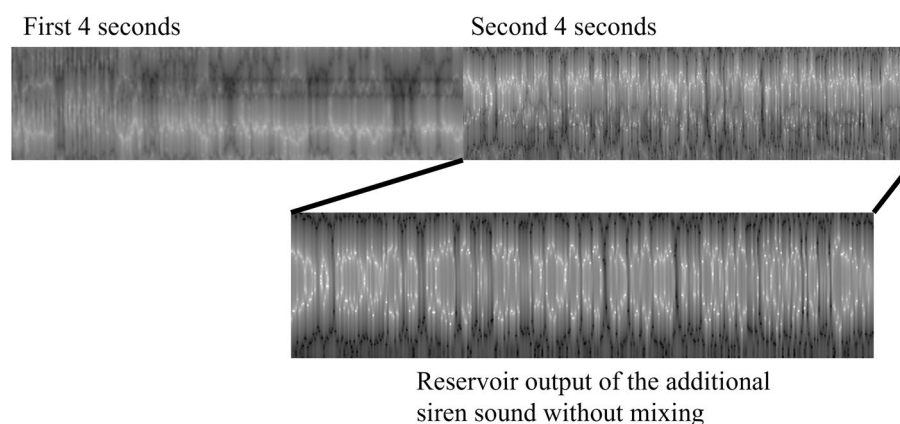


Figure 10. A noise resistance test using audio features generated from the urban sound recognition task. During the first four seconds of this eight second clip, drilling and car horn sounds are mixed, and the last four seconds contains the siren sound with a high amplitude (two times larger as compared to other two audio classes) is added to the mixed data. As shown in the figure, the latter four seconds of audio features shows high similarity compared to the reference siren sound.

with the same Hopf reservoir computer (an Euclidean distance less than 8). We anticipate a pattern matching algorithm originating from computer vision applications could be employed in this type of audio event separation and processing.

The implementation of this convolutional neural network adopts the same machine learning approach proposed by³⁰. Using the same urban sound recognition task, this allows a direct comparison of the features extracted from the physical reservoir computer as well as the spectrogram technique that is normally applied. Using the same machine learning readout but without computationally expensive preprocessing of the audio, the physical reservoir computing architecture employed in this paper achieved a 10% accuracy improvement compared to³⁰. In realistic applications for the Internet of Things, this machine learning method can be applied using dedicated neural processors, such as the Syntiant ND101. This particular chip could deploy approximately 60,000 neural cores, well above the requirement of the machine learning model used in the paper (~40,000 neural cores). As an alternative approach, the features generated from the reservoir computer could be further engineered to compress the amount of the data for audio recognition, such that the models can be deployed on low-level edge processors.

There are still limits in the reservoir computing method using the Hopf oscillator in its current form. First, the high accuracy sound event recognition requires many virtual nodes to generate diverse features for machine perception. However, increasing the virtual nodes leads to exponential growth of the sampling rate to read high quality audio data. We are actively seeking solutions to separate audio features from the original signal for recognition and recording, which could decrease the required sampling rate. Second, the current circuit-based physical reservoir separates the process of signal mixing and activation of the circuit. Redesigning the circuit is necessary to simplify signal reading for future system deployment. However, the ultimate version of the Hopf reservoir using MEMS will solve this problem, since the computing will happen on the audio sensing mechanisms. Lastly, the signal processing still relies on a digital readout. Though the algorithm is remarkably simple, a microcontroller unit is needed. We anticipate that the short-term solution will be deploying the optimized machine learning model as firmware (consuming less than 1 MB size of static memory without optimization and less than 256 KB dynamic memory for training upgraded machine learning models). A future goal should be using an analog circuit that could detect the spike signals for audio recognition (similar to neurons) to achieve a fully analog computer on edge devices⁴⁰.

Methods

The Hopf physical reservoir computer is realized through a proprietary circuit design proposed by²⁴. Following the schematic given in Fig. 11, the circuit is implemented using TL082 operational amplifiers and AD633 multipliers. The input audio signal is first normalized to the range from -1 to $+1$ and mixed with the sinusoidal forcing signal in MATLAB, then it is sent to the circuit by a National Instrument (NI) cDAQ-9174 data I/O module. The outputs from the circuit, referred to as the x and y states of the Hopf oscillator, are collected with a sampling rate of 10^5 samples/s by the same NI cDAQ-9174 for later machine learning processing.

Three datasets are employed in the sound recognition experiments. These consist of urban sound recognition, Qualcomm voice command, and spoken digits. The urban sound recognition dataset consists of 873 audio clips of 10 classes, which are high quality urban sound clips recorded in New York City⁴¹. Each audio clip is four seconds long with a sampling rate of at least 44.1 kHz. Compared to commonly available datasets, we have an extremely small number of samples.

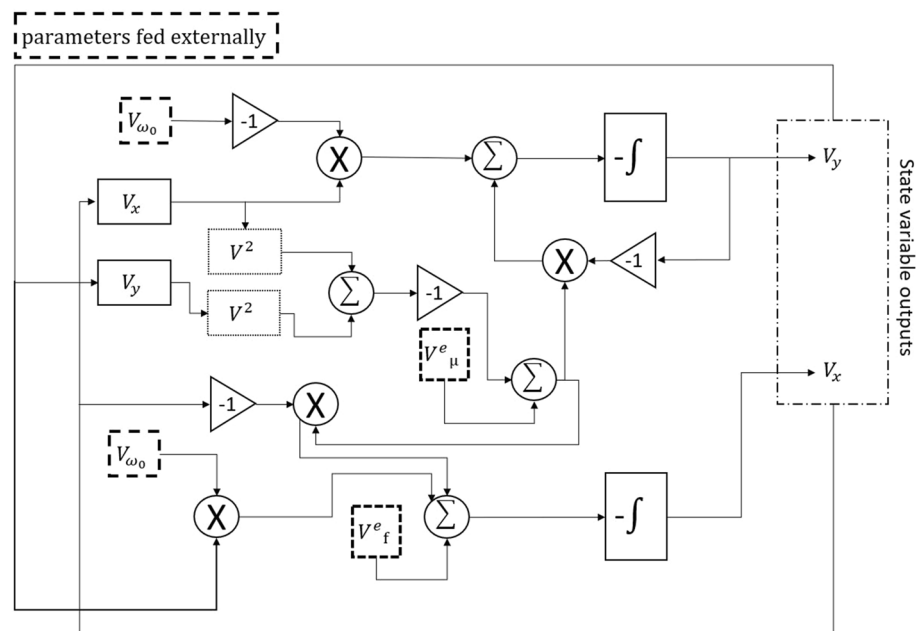


Figure 11. A simplified circuit schematic of the Hopf reservoir computer.

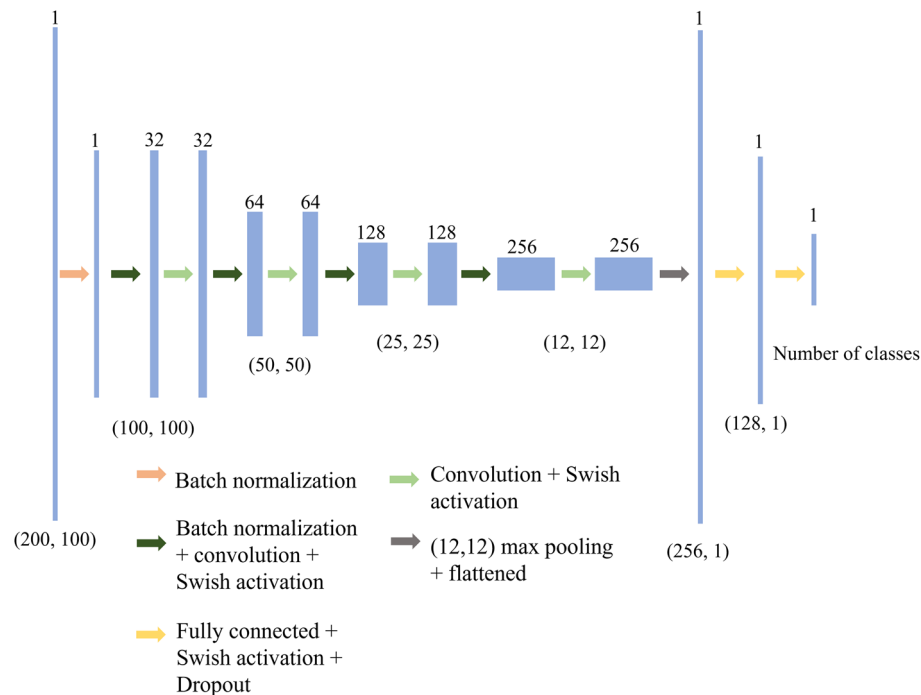


Figure 12. A schematic showing the convolutional neural network-based machine learning readout for classification of the audio events using the Hopf reservoir computer. The light blue boxes in the figure correspond to the feature maps generated from each machine learning operations. The arrows are the different machine learning operations. The numbers above the light blue boxes are the depth of feature maps, and the bottom numbers are the length and the width of the feature maps, respectively. A max pooling with a size of (2,2) is also operated after two consecutive convolutions to reduce the dimension of the feature maps. Note for the length and width, we only label the dimensions that are changed after machine learning operations.

To demonstrate reconfigurability of the Hopf reservoir computer for audio processing, the Qualcomm voice command dataset is also used. This dataset consists of 4270 audio clips with each clip lasting 1 second, which are four wake words that are collected from speakers with diverse speaking speeds and accents⁴². From the dataset, we use 1000 clips for experiments. Compared to the previous urban sound recognition case, the only difference in the processing algorithm is the retraining of the output portion (i.e., after convolution layers) of the machine learning readout (details are discussed in the later part of the methodology section and results section of the paper). To compare the proposed Hopf reservoir with other reservoirs, we also conduct an experiment of spoken digits recognition, which serves as the standard benchmarking test for reservoir computing. The spoken digits dataset consists of 3000 audio clips, which are spoken by five different speakers⁴³. As with the Qualcomm voice command dataset, the total number of audio clips for the experiments is set to be only 1000.

For the sake of processing speed, we resample each audio clip with a sampling rate of 4000 Hz and normalize the data to the range from -1 to $+1$ before sending to the analog circuit. 80% of the outputs from the circuit are used for training the machine learning model with the remaining 20% used for testing.

In Fig. 1, the nodal connections of the Hopf physical reservoir computer are shown. Although we only collect a 1D data stream from the Hopf circuit, the data stream consists of both input signals and the response from the virtual nodes defined by the sampling speed of the signals⁴⁴. We follow this principle of arranging and manipulating signals by their virtual nodes. The output from the circuit reservoir is first activated using an inverse hyperbolic tangent function^{24,45}.

$$x_{feature} = \tanh^{-1} x \quad (6)$$

Subsequently, the activated output is rearranged by the order of the virtual nodes as the feature maps for the machine perception. A sample feature map rendering consisting of 10 different classes of urban sound is shown as Fig. 2. The Hopf reservoir computer produces this feature map as described in “Hopf oscillator and reservoir” Section, which is then used as an input to the neural network shown in Figure 12. Effectively, the Hopf reservoir computer is offloading the costs of the computationally expensive Mel spectrum. A Swiss activation⁴⁶ is employed to boost the performance of the machine learning model on processing sparse neuron activation (i.e., dead neuron problems) and the overall accuracy of the machine learning model processing audio data. Note that a future version of the machine learning software using skipped connection (generating residual networks)⁴⁷ will further boost the robustness of the software for large set of data. Each 1 second clip of the outputs is further skip-sampled to a 200 (number of time samples) \times 100 (number of virtual nodes) for machine learning processing (as labeled in Fig. 12). The machine learning algorithm is implemented using Keras⁴⁸ with a TensorFlow backend. The training is conducted on an Nvidia RTX 2080Ti GPU and uses an Adam optimizer with the default

learning rate of 0.001⁴⁹. The loss function is cross entropy⁵⁰. The batch size during training is 5; the epochs is 100 for urban sound recognition dataset, 20 for Qualcomm voice command dataset, and 100 for the spoken digits.

Data availability

The datasets used and analysed during the current study available from the corresponding author on reasonable request.

Received: 20 February 2023; Accepted: 23 May 2023

Published online: 30 May 2023

References

- Lee, W. *et al.* Biosignal sensors and deep learning-based speech recognition: A review. *Sensors* **21**(4), 1399 (2021).
- Karmakar, P., Teng, S. W. & Lu, G. Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition. arXiv preprint [arXiv:2102.07259](https://arxiv.org/abs/2102.07259) (2021).
- Filho, C. P. *et al.* A systematic literature review on distributed machine learning in edge computing. *Sensors* **22**(7), 2665 (2022).
- Li, C. Openai's gpt-3 language model: A technical overview. *Blog Post* (2020).
- Patterson, D. *et al.* The carbon footprint of machine learning training will plateau, then shrink. *Computer* **55**(7), 18–28 (2022).
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLevey, C. & Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. <https://cdn.openai.com/papers/whisper.pdf> (2021). Accessed 28 Sept 2022.
- Adversa. The Road to Secure and Trusted AI. <https://adversa.ai/report-secure-and-trusted-ai/> (2021). Accessed 28 Sept 2022.
- IBM Security. Cost of a data breach 2022. <https://www.ibm.com/reports/data-breach> (2022). Accessed 28 Sept 2022.
- Garg, R. Open data privacy and security policy issues and its influence on embracing the internet of things. *First Monday* (2018).
- Deep, S. *et al.* A survey of security and privacy issues in the internet of things from the layered context. *Trans. Emerg. Telecommun. Technol.* **33**(6), e3935 (2022).
- Hao, K. Training a single AI model can emit as much carbon as five cars in their lifetimes (2019). <https://www.technologyreview.com/2019/06/06/239031> (2019). Accessed 28 Sept 2022.
- Fernando, C. & Sojakka, S. Pattern recognition in a bucket. In *European Conference on Artificial Life* 588–597 (Springer, 2003).
- Tanaka, G. *et al.* Recent advances in physical reservoir computing: A review. *Neural Netw.* **115**, 100–123 (2019).
- Shougat, M. R., Li, X., Mollik, T. & Perkins, E. An information theoretic study of a duffing oscillator array reservoir computer. *J. Comput. Nonlinear Dyn.* **16**(8), 081004 (2021).
- Morán, A. *et al.* Hardware-optimized reservoir computing system for edge intelligence applications. *Cogn. Comput.* <https://doi.org/10.1007/s12559-020-09798-2> (2021).
- Usami, Y. *et al.* In-materio reservoir computing in a sulfonated polyaniline network. *Adv. Mater.* **33**(48), 2102688 (2021).
- Moon, J. *et al.* Temporal data classification and forecasting using a memristor-based reservoir computing system. *Nat. Electron.* **2**(10), 480–487 (2019).
- Mizrahi, A. *et al.* Neural-like computing with populations of superparamagnetic basis functions. *Nat. Commun.* **9**(1), 1–11 (2018).
- Grolier, J. *et al.* Neuromorphic spintronics. *Nat. Electron.* **3**(7), 360–370 (2020).
- Larger, L. *et al.* High-speed photonic reservoir computing using a time-delay-based architecture: Million words per second classification. *Phys. Rev. X* **7**(1), 011015 (2017).
- Barazani, B., Dion, G., Morissette, J.-F., Beaudoin, L. & Sylvestre, J. Microfabricated neuroaccelerometer: Integrating sensing and reservoir computing in mems. *J. Microelectromech. Syst.* **29**(3), 338–347 (2020).
- Kan, S. *et al.* Simple reservoir computing capitalizing on the nonlinear response of materials: Theory and physical implementations. *Phys. Rev. Appl.* **15**(2), 024030 (2021).
- Appeltant, L. *et al.* Information processing using a single dynamical node as complex system. *Nat. Commun.* **2**(1), 1–6 (2011).
- Shougat, M. R. E. U., Li, X. F., Mollik, T. & Perkins, E. A Hopf physical reservoir computer. *Sci. Rep.* **11**(1), 1–13 (2021).
- Shougat, M. R. E. U., Li, X. F. & Perkins, E. Dynamic effects on reservoir computing with a Hopf oscillator. *Phys. Rev. E* **105**(4), 044212 (2022).
- Li, X. F. *et al.* Stochastic effects on a Hopf adaptive frequency oscillator. *J. Appl. Phys.* **129**(22), 224901 (2021).
- Li, X. F. *et al.* A four-state adaptive Hopf oscillator. *PLoS ONE* **16**(3), e0249131 (2021).
- Shougat, M. R., Kennedy, S. & Perkins, E. A self-sensing shape memory alloy actuator physical reservoir computer. *IEEE Sens. Lett.* <https://doi.org/10.1109/LSSENS.2023.3270704> (2023).
- Nayfeh, A. H. & Balachandran, B. *Applied Nonlinear Dynamics: Analytical, Computational, and Experimental Methods* (John Wiley & Sons, Hoboken, 2008).
- Yun, J., Srivastava, S., Roy, D., Stohs, N., Mydlarz, C., Salman, M., Steers, B., Bello, J. P. & Arora, A. Infrastructure-free, deep learned urban noise monitoring at 100mW. *CoRR* (2022).
- Gao, Y., Liu, Y., Zhang, H., Li, Z., Zhu, Y., Lin, H. & Yang, M. Estimating gpu memory consumption of deep learning models. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* 1342–1352 (2020).
- Lin, J., Zhu, L., Chen, W. M., Wang, W. C., Gan, C. & Han, S. On-device training under 256kb memory. arXiv preprint [arXiv:2206.15472](https://arxiv.org/abs/2206.15472) (2022).
- Rajaby, E. & Sayedi, S. M. A structured review of sparse fast Fourier transform algorithms. *Digit. Signal Process.* **123**, 103403 (2022).
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020).
- Kaur, A., Kaur, L. & Gupta, S. Image recognition using coefficient of correlation and structural similarity index in uncontrolled environment. *Int. J. Comput. Appl.* **59**(5) (2012).
- Sazli, M. H. A brief review of feed-forward neural networks. *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering* **50**(01) (2006).
- Wang, L., Zhang, Y. & Feng, J. On the Euclidean distance of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1334–1339 (2005).
- Lenk, C., Ekin, A., Rangelow, I. W. & Gutschmidt, S. Active, artificial hair cells for biomimetic sound detection based on active cantilever technology. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 4488–4491 (IEEE, 2018).
- Gomez, F., Lorimer, T. & Stoop, R. Signal-coupled subthreshold Hopf-type systems show a sharpened collective response. *Phys. Rev. Lett.* **116**, 108101 (2016).
- Ma, S., Brooks, D. & Wei, G.-Y. A binary-activation, multi-level weight RNN and training algorithm for ADC-/DAC-free and noise-resilient processing-in-memory inference with eNVM. arXiv preprint [arXiv:1912.00106](https://arxiv.org/abs/1912.00106) (2019).
- Salamon, J., Jacoby, C. & Bello, J. P. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia* 1041–1044 (2014).

42. Kim, B., Lee, M., Lee, J., Kim, Y. & Hwang, K. Query-by-example on-device keyword spotting. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* 532–538 (IEEE, 2019).
43. Jackson, Z. Free Spoken Digit Dataset (FSDD). <https://github.com/Jakobovski/free-spoken-digit-dataset> (2018). Accessed 28 Sept 2022.
44. Jacobson, P., Shirao, M., Kerry, Yu., Guan-Lin, S. & Ming, C. W. Hybrid convolutional optoelectronic reservoir computing for image recognition. *J. Lightwave Technol.* **40**(3), 692–699 (2021).
45. Miller, C. L. & Freedman, R. The activity of hippocampal interneurons and pyramidal cells during the response of the hippocampus to repeated auditory stimuli. *Neuroscience* **69**(2), 371–381 (1995).
46. Ramachandran, P., Zoph, B. & Le, Q. V. Searching for activation functions. arXiv preprint [arXiv:1710.05941](https://arxiv.org/abs/1710.05941) (2017).
47. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
48. Chollet, F. Keras: Deep Learning for humans. <https://github.com/keras-team/keras> (2015). Accessed 28 Sept 2022.
49. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014).
50. De Boer, P.-T., Kroese, D. P., Mannor, S. & Rubinstein, R. Y. A tutorial on the cross-entropy method. *Ann. Oper. Res.* **134**(1), 19–67 (2005).

Acknowledgements

The authors also greatly appreciate the fruitful discussion of the experimental procedures and results with Dr. Omar Zahr and Dr. Helge Seetzen.

Author contributions

M.R.E.U.S., X.L., S.S., K.W.M., and E.P. conceived the concepts and perspectives in this article together and co-wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023