Exploring neural oscillations during speech perception via surrogate gradient spiking neural networks

Alexandre Bittar[1,2,*] **and Philip N. Garner**[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Switzerland
[*]abittar@idiap.ch

ABSTRACT

Understanding cognitive processes in the brain demands sophisticated models capable of replicating neural dynamics at large scales. We present a physiologically inspired speech recognition architecture, compatible and scalable with deep learning frameworks, and demonstrate that end-to-end gradient descent training leads to the emergence of neural oscillations in the central spiking neural network. Significant cross-frequency couplings, indicative of these oscillations, are measured within and across network layers during speech processing, whereas no such interactions are observed when handling background noise inputs. Furthermore, our findings highlight the crucial inhibitory role of feedback mechanisms, such as spike frequency adaptation and recurrent connections, in regulating and synchronising neural activity to improve recognition performance. Overall, on top of developing our understanding of synchronisation phenomena notably observed in the human auditory pathway, our architecture exhibits dynamic and efficient information processing, with relevance to neuromorphic technology.

Introduction

In the field of speech processing technologies, the effectiveness of training deep artificial neural networks (ANNs) with gradient descent has led to the emergence of many successful encoder-decoder architectures for automatic speech recognition (ASR), typically trained in an end-to-end fashion over vast amounts of data[1–4]. Despite recent efforts[5–8] towards understanding how these ANN representations can compare with speech processing in the human brain, the cohesive integration of the fields of deep learning and neuroscience remains a challenge. Nonetheless, spiking neural networks (SNNs), a type of artificial neural network inspired by the biological neural networks in the brain, present an interesting convergence point of the two disciplines. Although slightly behind in terms of performance compared to ANNs, SNNs have recently achieved concrete progress[9] on speech command recognition tasks using the surrogate gradient method[10] which allows them to be trained via gradient descent. Further work has also shown that they can be used to define a spiking encoder inside a hybrid ANN-SNN end-to-end trainable architecture on the more challenging task of large vocabulary continuous speech recognition[11]. Their successful inclusion into contemporary deep learning ASR frameworks offers a promising path to bridge the existing gap between deep learning and neuroscience in the context of speech processing. This integration not only equips deep learning tools with the capacity to engage in speech neuroscience but also offers a scalable approach to simulate spiking neural dynamics, which supports the exploration and testing of hypotheses concerning the neural mechanisms and cognitive processes related to speech.

In neuroscience, various neuroimaging techniques such as electroencephalography (EEG) can detect rhythmic and synchronised postsynaptic potentials that arise from activated neuronal assemblies. These give rise to observable neural oscillations, commonly categorised into distinct frequency bands: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), low-gamma (30-80 Hz), and high-gamma (80-150 Hz)[12]. It is worth noting that while these frequency bands provide a useful framework, their boundaries are not rigidly defined and can vary across studies. Nevertheless, neural oscillations play a crucial role in coordinating brain activity and are implicated in cognitive processes such as attention[13–16], memory[17], sensory perception[18,19] and motor function[20,21]. Of particular interest is the phenomenon of cross-frequency coupling (CFC) which reflects the interaction between oscillations occurring in different frequency bands[14,22]. As reviewed by Abubaker *et al.*[23], many studies have demonstrated a relationship between CFC and working memory performance[24,25]. In particular phase-amplitude coupling (PAC) between theta and gamma rhythms appears to support memory integration[26–28], preservation of sequential order[29–31] and information retrieval[32]. In contrast, alpha-gamma coupling commonly manifests itself as a sensory suppression mechanism during selective attention[33,34], inhibiting task-irrelevant brain regions[35] and ensuring controlled access to stored knowledge[36]. Finally, beta oscillations are commonly associated with cognitive control and top-down processing[37].

In the context of speech perception, numerous investigations have revealed a similar oscillatory hierarchy, where the temporal organisation of high-frequency signal amplitudes in the gamma range is orchestrated by low-frequency neural phase dynamics, specifically in the delta and theta ranges[38–42]. These three temporal scales – delta, theta and gamma – naturally manifest in speech and represent specific perceptual units. In particular, delta-range modulation (1-2 Hz) corresponds to perceptual groupings formed by lexical and phrasal units, encapsulating features such as the intonation contour of an utterance. Modulation within the theta-range aligns with the syllabic rate (4 Hz) around which the acoustic envelope consistently oscillates. Finally, (sub)phonemic attributes, including formant transitions that define the fine structure of speech signals, correlate with higher modulation frequencies (30-50 Hz) within the low-gamma range. The close correspondence between the perception of (sub)phonemic, syllabic and phrasal attributes on one hand, and the manifestation of gamma, theta and delta neural oscillations on the other, was notably emphasised by Giraud and Poeppel[40]. These different levels of temporal granularity inherent to speech signals therefore appear to be processed in a hierarchical fashion, with the intonation and syllabic contour encoded by earlier neurons guiding the encoding of phonemic features by later neurons. Some insights about how phoneme features end up being encoded in the temporal gyrus were given by Mesgarani *et al.*[43]. Drawing from recent research[44] on the neural oscillatory patterns associated with the sentence superiority effect, it is suggested that such low-frequency modulation may facilitate automatic linguistic chunking by grouping higher-order features into packets over time, thereby contributing to enhanced sentence retention. The engagement of working memory in manipulating phonological information enables the sequential retention and processing of speech sounds for coherent word and sentence representations. Additionally, alpha modulation has also been shown to play a role in improving auditory selective attention[45–47], reflecting the listener's sensitivity to acoustic features and their ability to comprehend speech[48].

Computational models[41,49] have shown that theta oscillations can indeed parse speech into syllables and provide a reliable reference time frame to improve gamma-based decoding of continuous speech. These approaches[41,49] implement specific models for theta and gamma neurons along with a distinction between inhibitory and excitatory neurons. The resulting networks are then optimised to detect and classify syllables with very limited numbers of trainable parameters (10-20). In contrast, this work proposes to utilise significantly larger end-to-end trainable multi-layered architectures (400k-20M trainable parameters) where all neuron parameters and synaptic connections are optimised to predict sequences of phoneme/subword probabilities, that can subsequently be decoded into words. By avoiding constraints on theta or gamma activity, the approach allows us to explore which forms of CFC naturally arise when solely optimising the decoding performance. Even though the learning mechanism is not biologically plausible, we expect that a model with sufficiently realistic neuronal dynamics and satisfying ASR performance should reveal similarities with the human brain. We divide our analysis in two parts,

1. **Architecture**: As a preliminary analysis, we conduct hyperparameter tuning to optimise the model's architectural parameters. On top of assessing the network's capabilities and scalability, we notably evaluate how the incorporation of spike-frequency adaptation (SFA) and recurrent connections impact the speech recognition performance.

2. **Oscillations**: Here we delve into the core part of our analysis on the emergence of neural oscillations within our model. Each SNN layer is treated as a distinct neuron population, from which spike trains are aggregated into a population signal similar to EEG data. Through intra- and inter-layer CFC analysis, we investigate the presence of significant delta-gamma, theta-gamma, alpha-gamma and beta-gamma PAC. We also examine the impact of SFA and recurrent connections on the synchronisation of neural activity.

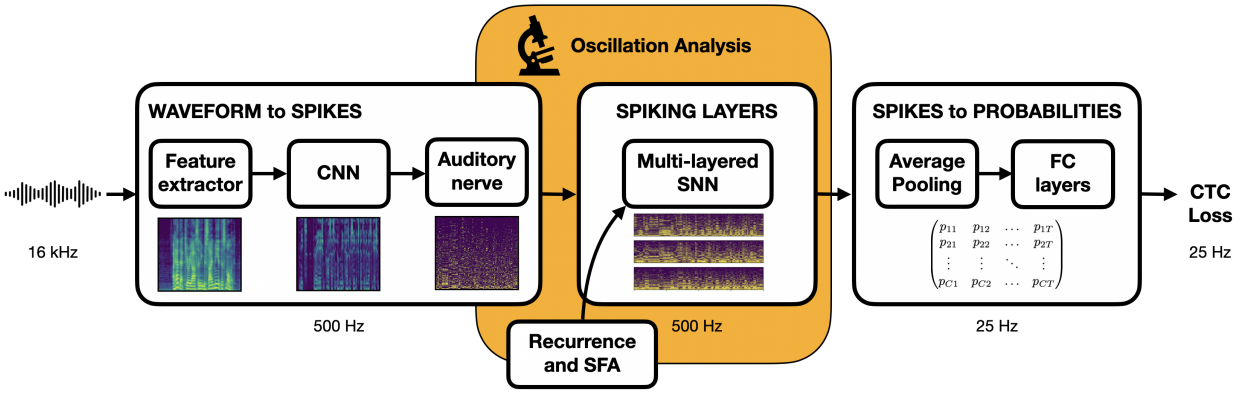

Figure 1. End-to-end trainable speech recognition pipeline. Input waveform is converted to a spike train representation to be processed by the central SNN before being transformed into output phoneme probabilities sent to a loss function for training.

Results

Architectural analysis

In order to draw a comparison with the human auditory pathway, we introduce a physiologically inspired ASR pipeline illustrated in Fig. 1. The proposed hybrid ANN-SNN architecture is trained in an end-to-end fashion on the TIMIT dataset[50] to predict phoneme sequences from speech waveforms. In the architectural design, we aimed to minimise the complexity of ANN components and favour the central SNN which will be the focus of the oscillation analysis. Here as a preliminary step, we examine how relevant hyperparameters affect the phoneme error rate (PER). On top of assessing the scalability of our approach to larger networks, we identify the importance of the interplay between recurrence and SFA.

Network scalability

As reported in Table 1, performance improves with added layers, peaking at 4-6 layers before declining, which suggests a significant contribution to the final representation from all layers within this range. Compared to conventional non-spiking recurrent neural network (RNN) encoders used in ASR, our results support the scalability of surrogate gradient SNNs to relatively deep architectures. Additionally, augmenting the number of neurons until about 1,000 per layer consistently yields lower PERs, beyond which performance saturates.

Table 1. Hyperparameter tuning for the number of SNN layers and neurons per layer on the TIMIT dataset. The second column gives both the number of trainable parameters in the multi-layered SNN (left) and in the whole encoder (right). The PERs are reported after 50 training epochs using a 5 ms time step, 16 convolutional neural network (CNN) channels, 50% of adaptive leaky integrate-and-fire (AdLIF) neurons, 100% feedforward and 50% recurrent connectivity. The performance of the architecture when removing the SNN is also reported (bottom). Bold values indicate the lowest achieved PERs.

Number of layers	Number of neurons per layer	Number of parameters	Test PER [%]	Validation PER [%]
1	512	740k / 1.3M	23.3	21.8
2	512	1.1M / 1.7M	21.0	19.2
3	512	1.5M / 2.1M	20.5	18.2
4	512	1.9M / 2.5M	20.2	**17.4**
5	512	2.3M / 2.9M	**20.0**	17.6
6	512	2.7M / 3.3M	**20.0**	17.9
7	512	3.1M / 3.7M	20.5	18.0
3	64	91k / 394k	30.9	29.6
3	128	211k / 547k	25.5	24.1
3	256	537k / 938k	22.5	20.9
3	768	3.0M / 3.7M	19.6	17.4
3	1024	4.9M / 5.7M	19.1	**17.1**
3	1536	10.1M / 11.2M	**19.0**	17.3
3	2048	17.1M / 18.5M	19.2	17.2
no nerve, no SNN		0 / 873k	34.2	32.0

Recurrent connections and spike-frequency adaptation

The impact of adding SFA in the neuron model as well as using recurrent connections are reported in Table 2. Interestingly, we find that without SFA, optimal performance is achieved by limiting the recurrent connectivity to 80%. When additionally using SFA, further limitation of the recurrent connectivity about 50% yields the lowest PER. This differs from conventional non-spiking RNNs, where employing fully-connected (FC) recurrent matrices is favoured. These results indicate that while requiring fewer parameters, an architecture with sparser recurrent connectivity and more selective parameter usage can achieve better task performance. Overall, SFA and recurrent connections individually yield significant error rate reduction, although they respectively grow as $\mathcal{O}(N)$ and $\mathcal{O}(N^2)$ with the number of neurons N. In line with previous studies on speech command recognition tasks[9,51], our results emphasise the metabolic and computational efficiency gained by harnessing the heterogeneity of adaptive spiking neurons. Furthermore, effectively calibrating the interplay between unit-wise feedback from SFA and layer-wise feedback from recurrent connections appears crucial for achieving optimal performance.

Table 2. Ablation experiments for the recurrent connectivity and proportion of neurons with SFA on the TIMIT dataset. PERs are reported after 50 epochs using a 5 ms time step, 16 CNN channels, 3 layers, 512 neurons per layer and 100% feedforward connectivity. Bold values indicate the lowest achieved PERs.

Model type	Recurrent connectivity	Proportion of AdLIF neurons	Number of parameters	Test PER [%]	Validation PER [%]
No recurrence no SFA	0	0	1.1M / 1.7M	26.9	24.8
Recurrence only	0.2	0	1.3M / 1.8M	22.0	20.1
	0.5	0	1.5M / 2.1M	21.0	18.9
	0.8	0	1.8M / 2.3M	20.8	18.7
	1	0	1.9M / 2.5M	21.8	19.3
SFA only	0	0.2	1.1M / 1.7M	24.2	21.7
	0	0.5	1.1M / 1.7M	23.7	21.6
	0	0.8	1.1M / 1.7M	23.3	21.0
	0	1	1.1M / 1.7M	22.9	21.1
Recurrence and SFA	0.2	0.2	1.3M / 1.8M	20.9	19.3
	0.5	0.5	1.5M / 2.1M	**20.5**	**18.2**
	0.8	0.8	1.8M / 2.3M	21.2	18.8
	1	1	1.9M / 2.5M	23.3	21.5

Corrolary

We observe in Supplementary Table 1 that decreasing the simulation time step does not affect the performance. Although making the simulation of spiking dynamics more realistic, one might have anticipated that backpropagating through more time steps could hinder the training and worsen the performance as observed in standard RNNs often suffering from vanishing or exploding gradients[52]. With inputs ranging from 1,000 to over 7,000 steps using 1 ms intervals on TIMIT, our results demonstrate a promising scalability of surrogate gradient SNNs for processing longer sequences. Secondly, as reported in Supplementary Table 2, we did not observe substantial improvement when increasing the number of auditory nerve fibers past $\sim$5,000, even though there are approximately 30,000 of them in the human auditory system. This could be due to both the absence of a proper model for cochlear and hair cell processing in our pipeline and to the relatively low number of neurons (<1,000) in the subsequent layer. Finally, as detailed in Supplementary Table 3, reduced feedforward connectivity in the SNN led to poorer overall performance. This contrasts with our earlier findings on recurrent connectivity, highlighing the distinct functional roles of feedforward and feedback mechanisms in the network.

Oscillation analysis

Based on our previous architectural results that achieved satisfactory speech recognition performance using a physiologically inspired model, we hypothesise that the spiking dynamics of a trained network should, to some extent, replicate those occurring throughout the auditory pathway. Our investigation aims to discern if synchronisation phenomena resembling brain rhythms manifest within the implemented SNN as it processes speech utterances to recognise phonemes.

Synchronised gamma activity produces low-frequency rhythms

As illustrated in Fig. 2a, the spike trains produced by passing a test-set speech utterance through the trained architecture exhibit distinct low-frequency rhythmic features in all layers. By looking at the histogram of single neuron firing rates illustrated in Fig. 2b, we observe that the distribution predominantly peaks at gamma range, with little to no activity below beta. This reveals that the low-frequency oscillations visible in Fig. 2a actually emerge from the synchronisation of gamma-active neurons, which is understood to be a key operational feature of activated cortical networks. The resulting low-frequency rhythms appear to follow to some degree the intonation and syllabic contours of the input filterbank features and to persist across layers. Compared to the three subsequent layers, higher activity in the auditory nerve comes from the absence of inhibitory SFA and recurrence mechanisms. These initial observations suggest that the representation of higher-order features in the last layer is temporally modulated by lower level features already encoded in the auditory nerve fibers, even though each layer is seen to exhibit distinct rhythmic patterns. In the next section, we focus on measuring this modulation more rigorously via PAC analysis.

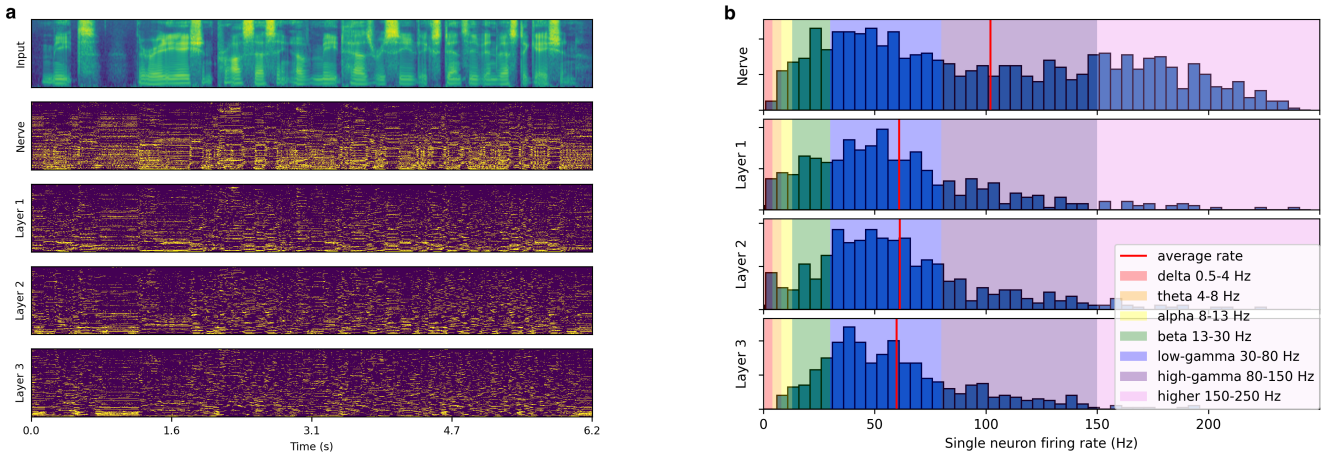

Figure 2. Spiking activity in response to speech input. **a** Input filterbank features and resulting spike trains produced across layers. For each layer, the neurons are vertically sorted on the y-axis by increasing average firing rate (top to bottom). The model uses a 2 ms time step, 16 CNN channels, 3 layers of size 512, 50% AdLIF neurons, 100% feedforward and 50% recurrent connectivity. **b** Corresponding distribution of single neuron firing rates.

Phase-amplitude coupling within and across layers

By aggregating over the relevant spike trains, we compute distinct EEG-like population signals for the auditory nerve fibers and each of the three SNN layers. These are then filtered in the different frequency bands, as illustrated in Fig. 3a, which allows us to perform CFC analyses. We measure PAC within-layer and across-layers between all possible combinations of frequency bands and find multiple significant forms of coupling for every utterance. An example of significant theta low-gamma coupling between the auditory nerve fibers and the last layer is illustrated in Fig. 3b. Here input low-frequency modulation is observed to significantly modulate the output gamma activity. The network therefore integrates and propagates intonation and syllabic contours across layers via a synchronised neural activity along these perceptual cues. It is important to note that the synchronisation of neural signals to the auditory envelope emerged without imposing any theta or gamma activity in our network. The optimisation of the PER combined with sufficiently realistic spiking neuronal dynamics therefore represent sufficient conditions to reproduce some broad properties of neural oscillations observed in the brain, suggesting a general functional role of facilitating information processing. More generally, the patterns of coupling between different neural populations and frequency bands were found to differ from one utterance to another. These variations indicate that the neural processing of our network is highly dynamic and depends on the acoustic properties and linguistic content of the input. The observed rich range of intra- and inter-layer couplings suggests that the propagation of low-level input features such as intonation and syllabic rhythm is only one part of these synchronised neural dynamics.

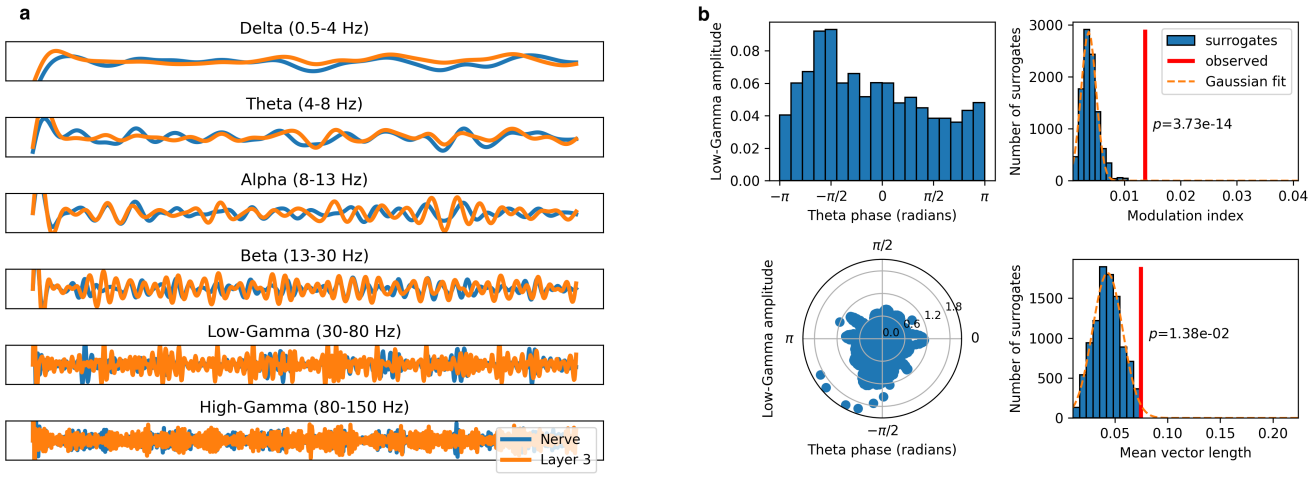

Figure 3. Cross-frequency coupling of population aggregated activity. **a** Population signals of auditory nerve fibers (blue) and last layer (orange) filtered in distinct frequency bands. **b** Modulation index and mean vector length metrics as measures of PAC between the theta band of the auditory nerve fibers and the low-gamma band of the last layer.

Impact of spike-frequency adaptation and recurrent connections on oscillations

As presented in Table 3, we record an overall average activity of 100 Hz in a trained architecture with no SFA and no recurrent connections. Using recurrent connections but no SFA, the average activity drops to 71 Hz where 60% of trained recurrent weights are inhibitory (see Supplementary Fig. 2) and help regulate the activity. Using SFA with or without recurrent connections further decreases the firing rate to 61 Hz emphasising the overall inhibitory effect of SFA. This regularisation of the network activity is able to improve the encoding of speech through the architecture as it results in better ASR performance. While both forms of feedback exhibit an overall inhibitory effect, SFA operates at the individual neuron level whereas recurrent connections act at the layer level. Our analysis indicates distinct influences of these two mechanisms on CFCs. On the one hand, inhibitory feedback produced by SFA reinforces synchronisation rhythms in both intra- and inter-layer interactions, especially within theta and alpha bands. This is in line with the work of Crook *et al.*[53], which demonstrated that SFA can encourage and stabilise the synchronisation of cortical networks. Similarly, Augustin *et al.*[54] showed that adaptation promotes periodic signal propagation. In our case, the resulting temporal reorganisation and activity regulation occurring at the neuron level appear to enable more effective parsing and encoding of speech information, leading to improved ASR performance. It is interesting to note that the adaptation time constant $\tau_w \in [30, 350]$ ms corresponds to delta-beta ranges, whereas the membrane potential time constant $\tau_u \in [3, 25]$ ms corresponds to gamma range. SFA therefore naturally enables the production of gamma bursts occurring at lower frequencies. On the other hand, networks with recurrent connections are associated with an increase in intra-layer couplings especially in the last layer, yet a decrease in inter-layer couplings. Layer-wise recurrence therefore enables more complex intra-layer dynamics but reduces inter-layer synchronisation, resulting in more localised rhythms.

Table 3. Comparison of the oscillatory activity resulting from passing the 64 longest TIMIT test-set utterances through networks with and without SFA and recurrent connections. The first column reports the average firing rate of each network. Columns delta to beta then give the number of significant PACs measured for each modulating frequency band across all intra- and inter-layer interactions. The last column gives the total number of significant PACs across all frequency bands. All networks use a 2 ms time step, 16 CNN channels, 3 layers, 512 neurons per layer and 100% feedforward connectivity.

Model Type	**Firing rate** [Hz]	**Delta** (intra, inter)	**Theta** (intra, inter)	**Alpha** (intra, inter)	**Beta** (intra, inter)	**Total** (intra, inter)
No recurrence no SFA	100	(11, 21)	(39, 73)	(61, 69)	(22, 29)	(133, 192)
Recurrence only	71	(12, 3)	(67, 39)	(62, 61)	(39, 29)	(180, 132)
SFA only	61	(14, 29)	(53, 161)	(102, 162)	(44, 38)	(213, 390)
Recurrence and SFA	61	(12, 13)	(64, 51)	(103, 79)	(86, 34)	(265, 177)

Table 4 outlines the five most prominent coupling patterns observed within each architecture. The activity of the last SNN layer stands out as the most significantly modulated overall. By architectural construction, modulation in that final layer has the most impact on the ASR task as the corresponding spike trains are the ones converted to phoneme probabilities using a small ANN. In the last layer, an increase in the occurrences of theta, alpha and beta intra-layer couplings goes together with a decrease of the PER, which suggests more selective integration of phonetic features, enhanced attentional processes, as well as improved assimilation of contextual information.

Table 4. Most measured coupling types. After analysing the 64 longest utterances of TIMIT test-set, we report the five types of coupling for which significant PAC was measured on most utterances. Models with and without SFA and recurrent connections are compared.

Model type	Number of Utterances	Phase Signal	Amplitude Signal	Low-freq Band	High-freq Band
No Recurrence no SFA	32	Nerve	First Layer	Theta	Low-Gamma
	28	Nerve	Nerve	Theta	Low-Gamma
	26	Nerve	Nerve	Alpha	Low-Gamma
	21	Nerve	First Layer	Alpha	Low-Gamma
	12	Nerve	Third Layer	Alpha	Low-Gamma
Recurrence only	31	Third Layer	Third Layer	Alpha	Low-Gamma
	28	Nerve	Nerve	Theta	Low-Gamma
	21	Third Layer	Third Layer	Theta	Low-Gamma
	20	Nerve	Third Layer	Alpha	Low-Gamma
	20	Third Layer	Third Layer	Beta	Low-Gamma
SFA only	51	Third Layer	Third Layer	Alpha	Low-Gamma
	42	Second Layer	Third Layer	Theta	High-Gamma
	36	Nerve	Third Layer	Alpha	Low-Gamma
	33	Second Layer	Third Layer	Alpha	Low-Gamma
	26	First Layer	Third Layer	Theta	High-Gamma
Recurrence and SFA	63	Third Layer	Third Layer	Alpha	Low-Gamma
	51	Third Layer	Third Layer	Beta	Low-Gamma
	27	Third Layer	Third Layer	Theta	Low-Gamma
	25	Nerve	Nerve	Theta	Low-Gamma
	19	Nerve	Third Layer	Alpha	Low-Gamma

Effects of training and input type on neuronal activity

In order to further understand the emergence of coupled signals, we consider passing speech through an untrained network, as well as passing different types of noise inputs through a trained network. Trained architectures exhibit persisting neuronal activity across layers compared to untrained ones, where the activity almost completely decays after the first layer. In trained networks, noise inputs lead to single neuron firing rate distributions peaking at very low rates and where the activity gradually decreases across layers, as illustrated in Fig. 4b. This contrasts with the response to speech inputs seen in Fig. 2b where the activity was sustained across layers with most of the distribution in the gamma range. We tested uniform noise as well as different noise sources (air conditioner, babble, copy machine and typing) from the the MS-SNSD dataset[55]. Compared to a speech input, all noise types yielded reduced average firing rates (from 60 Hz to about 40 Hz) with most of the neurons remaining silent. This highly dynamic processing of information is naturally efficient at attenuating its activity when processing noise or any input that does not induce sufficient synchronisation. Interestingly, babble noises were found in certain cases to induce some significant PAC patterns, whereas other noise types resulted in no coupling at all. Even though babble noises resemble speech and produced some form of synchronisation, they only triggered a few neurons per layer. Overall, we showed that synchronicity of neural oscillations in the form of PAC results from training and is only triggered when passing an appropriate speech input.

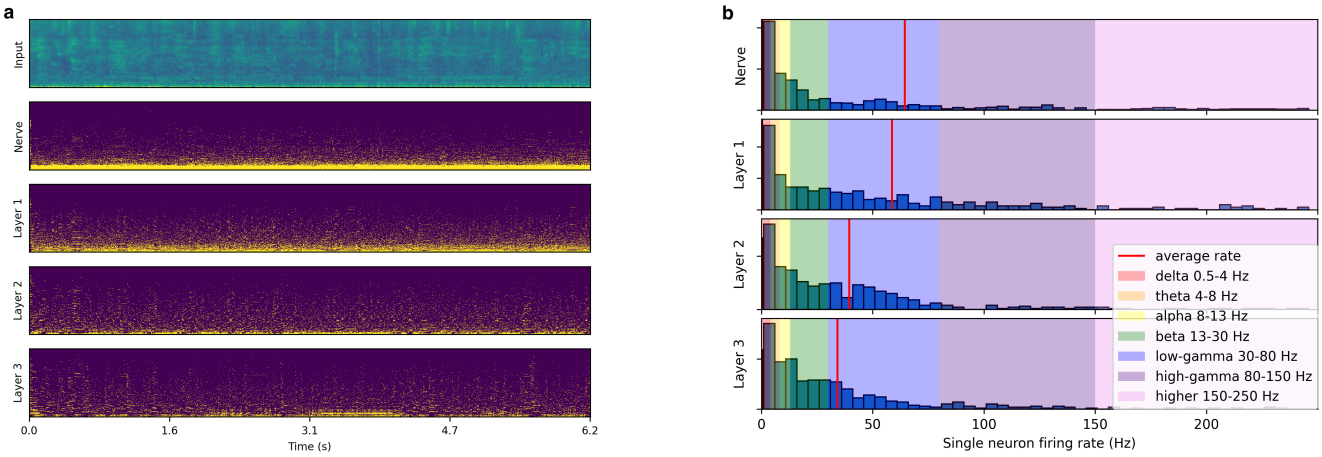

Figure 4. Spiking activity in response to babble noise input. **a** Input filterbank features and resulting spike trains produced across layers. The model uses a 2 ms time step, 16 CNN channels, 3 layers of size 512, 50% AdLIF neurons, 100% feedforward and 50% recurrent connectivity. **b** Corresponding single neuron firing rate distributions.

Scaling to a larger dataset

Our approach was extended to the Librispeech dataset[56] with 960 hours of training data. After 8 epochs, we reached 9.5% word error rate on the *test-clean* data split. As observed on TIMIT, trained models demonstrated similar CFCs in their spiking activity.

Training on speech command recognition task

With our experimental setup, the encoder is directly trained to recognise phonemes on TIMIT and subwords on Librispeech. One could therefore assume that the coupled gamma activity emerges from that constraint. In order to test this hypothesis, we run additional experiments on a speech command recognition task where no phoneme or subword recognition is imposed by the training. Instead the model is directly trained to recognise a set of short words. We use the same architecture as on TIMIT, except the average pooling layer is replaced by a readout layer as defined in Bittar and Garner[9] which reduces the temporal dimension altogether, as required by the speech command recognition task. Interestingly, using speech command classes as ground truths still produces significant PAC patterns, especially in the last layer. These results indicate that the emergence of the studied rhythms does not require phoneme-based training and may be naturally emerging from speech processing. Using the second version of the Google Speech Commands data set[57] with 35 classes, we achieve a test set accuracy of 96%, which, to the best of our knowledge, marginally improves upon the current state-of-the-art performance using SNNs of 95.35%[58].

Discussion

In this study, we introduced a physiologically inspired speech recognition architecture, centred around an SNN, and designed to be compatible with modern deep learning frameworks. As set out in the introduction, we first explored the capabilities and scalability of the proposed speech recognition architecture before analysing neural oscillations.

Our preliminary architectural analysis demonstrated a satisfactory level of scalability to deeper and wider networks, as well as to longer sequences and larger datasets. This scalability was achieved through our approach of utilising the surrogate gradient method to incorporate an SNN into an end-to-end trainable speech recognition pipeline. In addition, our ablation experiments emphasised the importance of including SFA within the neuron model, along with layer-wise recurrent connections, to attain optimal recognition performance.

The subsequent analysis of the spiking activity across our trained networks in response to speech stimuli revealed that neural oscillations, commonly associated with various cognitive processes in the brain, did emerge from training an architecture to recognise words or phonemes. Through CFC analyses, we measured similar synchronisation phenomena to those observed throughout the human auditory pathway. During speech processing, trained networks exhibited several forms of PAC, including delta-gamma, theta-gamma, alpha-gamma, and beta-gamma, while no such coupling occurred when processing pure background noise. Our networks' ability to synchronise oscillatory activity in the last layer was also associated with improved speech recognition performance, which points to a functional role for neural oscillations in auditory processing. Even though we employ gradient descent training, which does not represent a biologically plausible learning algorithm, our approach was capable of replicating natural phenomena of macro-scale neural coordination. By leveraging the scalability offered by deep learning frameworks, our approach can therefore serve as a valuable tool for studying the emergence and role of brain rhythms.

Building upon the main outcome of replicating neural oscillations, our analysis on SFA and recurrent connections emphasised their key role in actively shaping neural responses and driving synchronisation via inhibition in support of efficient auditory information processing. Our results point towards further work on investigating more realistic feedback mechanisms including efferent pathways across layers. More accurate neuron populations could also be obtained using clustering algorithms.

Aside from the fundamental aspect of developing the understanding of biological processes, our research on SNNs also holds significance for the fields of neuromorphic computing and energy efficient technology. Our exploration of the spiking mechanisms that drive dynamic and efficient information processing in the brain is particularly relevant for low-power audio and speech processing applications, such as on-device keyword spotting. In particular, the absence of synchronisation in our architecture when handling background noise results in fewer computations, making our approach well-suited for always-on models.

Methods

Spiking neuron model

Physiologically grounded neuron models such as the well known Hodgkin and Huxley model[59] can be reduced to just two variables[60,61]. More contemporary models, such as the Izhikevich[62] and adaptive exponential integrate-and-fire[63] models, have similarly demonstrated the capacity to accurately replicate voltage traces observed in biological neurons using just membrane potential and adaptation current as essential variables[64]. With the objective of incorporating realistic neuronal dynamics into large-scale neural network simulations with gradient descent training, the linear AdLIF neuron model stands out as an adequate compromise between physiological plausibility and computational efficiency. It can be described in continuous time by the following differential equations,

$$\tau_u \dot{u}(t) = -\big(u(t) - u_{\text{rest}}\big) - R\,w(t) + R\,I(t) - \tau_u\,(\vartheta - u_r) \sum_f \delta(t - t^f) \tag{1}$$

$$\tau_w \dot{w}(t) = -w(t) + a\big(u(t) - u_{\text{rest}}\big) + \tau_w\, b \sum_f \delta(t - t^f)\,. \tag{2}$$

The neuron's internal state is characterised by the membrane potential $u(t)$ which linearly integrates stimuli $I(t)$ and gradually decays back to a resting value u_{rest} with time constant $\tau_u \in [3, 25]$ ms. A spike is emitted when the threshold value ϑ is attained, $u(t) \geq \vartheta$, denoting the firing time $t = t^f$, after which the potential is reset by a fixed amount $\vartheta - u_r$. In the following, we will use $u_r = u_{\text{rest}}$ for simplicity. The second variable $w(t)$ is coupled to the potential with strength a and decay constant $\tau_w \in [30, 350]$ ms, characterising sub-threshold adaptation. Additionally, $w(t)$ experiences an increase of b after a spike is emitted, which defines spike-triggered adaptation. The differential equations can be simplified as,

$$\tau_u \dot{u}(t) = -u(t) - w(t) + I(t) - \tau_u \sum_f \delta(t - t^f) \tag{3}$$

$$\tau_w \dot{w}(t) = -w(t) + a\,u(t) + \tau_w\, b \sum_f \delta(t - t^f)\,. \tag{4}$$

by making all time-dependent quantities dimensionless with changes of variables,

$$u \to \frac{u - u_{\text{rest}}}{\vartheta - u_{\text{rest}}}\,, \quad w \to \frac{Rw}{\vartheta - u_{\text{rest}}} \quad \text{and} \quad I \to \frac{RI}{\vartheta - u_{\text{rest}}}\,,$$

and redefining neuron parameters as,

$$a \to Ra\,, \quad b \to \frac{Rb}{\vartheta - u_{\text{rest}}}\,, \quad \vartheta \to \frac{\vartheta - u_{\text{rest}}}{\vartheta - u_{\text{rest}}} = 1 \quad \text{and} \quad u_{\text{rest}} \to \frac{u_{\text{rest}} - u_{\text{rest}}}{\vartheta - u_{\text{rest}}} = 0\,.$$

This procedure gets rid of unnecessary parameters such as the resistance R, as well as resting, reset and threshold values, so that a neuron ends up being fully characterised by four parameters: τ_u, τ_w, a and b. As derived in the Supplementary Appendix, the differential equations can be solved in discrete time with step size Δt using a forward-Euler first-order exponential integrator method. After initialising $u_0 = w_0 = s_0 = 0$, and defining $\alpha := \exp\frac{-\Delta t}{\tau_u}$ and $\beta := \exp\frac{-\Delta t}{\tau_w}$, the neuronal dynamics can be solved by looping over time steps $t = 1, 2, \ldots, T$ as,

$$u_t = \alpha\big(u_{t-1} - s_{t-1}\big) + (1 - \alpha)\big(I_t - w_{t-1}\big) \tag{5}$$

$$w_t = \beta\big(w_{t-1} + b\,s_{t-1}\big) + (1 - \beta)\,a\,u_{t-1} \tag{6}$$

$$s_t = \big(u_t \geq 1\big)\,. \tag{7}$$

Stability conditions for the value of the coupling strength a are derived in the Supplementary Appendix. Additionally, we constrain the values of the neuron parameters to biologically plausible ranges[54,65],

$$\tau_u \in [3, 25]\,\text{ms}, \quad \tau_w \in [30, 350]\,\text{ms}, \quad a \in [-0.5, 5], \quad b \in [0, 2]\,. \tag{8}$$

This AdLIF neuron model is equivalent to a Spike Response Model (SRM) with continuous time kernel functions illustrated in Fig. 5. The four neuron parameters τ_u, τ_w, a and b all characterise the shape of these two curves. A derivation of the kernel-based SRM formulation is presented in the Supplementary Appendix.

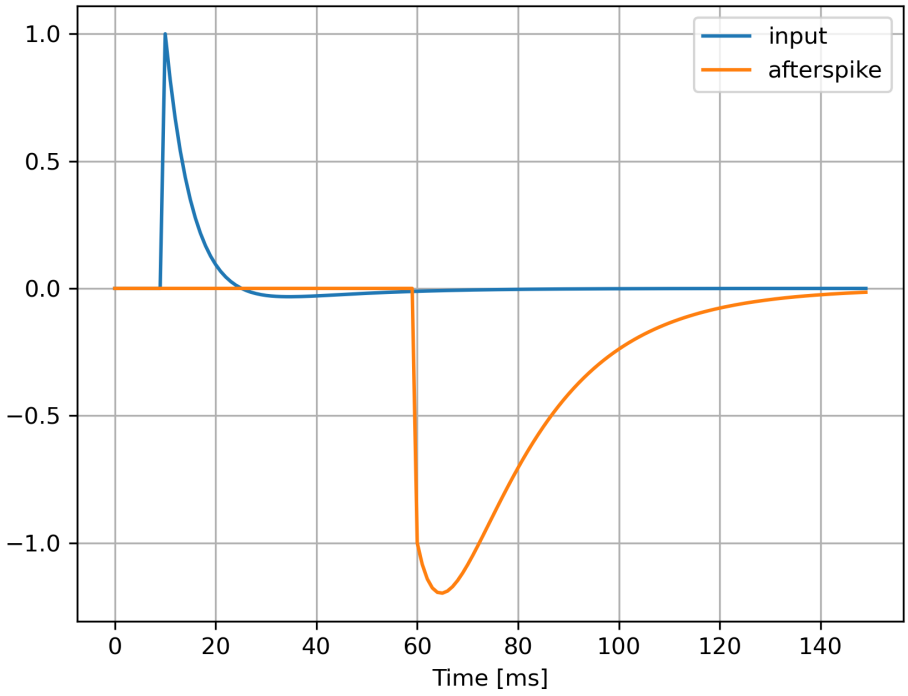

Figure 5. Kernel functions of AdLIF neuron model. Membrane potential response to an input pulse at $t = 10$ ms (in blue) and to an emitted spike at $t = 60$ ms (in orange). The neuron parameters are $\tau_u = 5$ ms, $\tau_w = 30$ ms, $a = 0.5$ and $b = 1.5$.

Spiking layers

The spiking dynamics described in Eqs. (5) to (7) can be computed in parallel for a layer of neurons. In a multi-layered network, the l-th layer receives as inputs a linear combination of the previous layer's outputs $s^{l-1} \in \{0,1\}^{B\times T\times N^{l-1}}$ where B, T and N^{l-1} represent the batch size, number of time steps and number of neurons respectively. Feedback from the l-th layer can also be implemented as a linear combination of its own outputs at the previous time step $s^l_{t-1} \in \{0,1\}^{N^l}$, so that the overall neuron stimulus I^l_t for neurons in the l-th layer at time step t is computed as,

$$I^l_t = W^l\, s^{l-1}_t + V^l\, s^l_{t-1}\,. \tag{9}$$

Here the feedforward $W^l \in \mathbb{R}^{N^{l-1}\times N^l}$ and feedback connections $V^l \in \mathbb{R}^{N^l\times N^l}$ are trainable parameters. Diagonal elements of V^l are set to zero as afterspike self inhibition is already taken into account in Eq. (5). Additionally, a binary mask can be applied to matrices W^l and V^l to limit the number of nonzero connections. Similarly, a portion of neurons in a layer can be reduced to leaky integrate-and-fire (LIF) dynamics without any SFA by applying another binary mask to the neuron adaptation parameters a and b. Indeed, if $a = b = 0$, the adaptation current vanishes $w_t = 0\ \forall t \in \{1,2,\dots,T\}$ and has no more impact on the membrane potential.

Surrogate gradient method

The only non-differentiable component of the derived neuronal dynamics lies in the threshold operation described in Eq. (7). To address this, a surrogate derivative can be manually specified using PyTorch[66], which enables the application of the Back-Propagation Through Time algorithm for training the resulting SNN in a manner similar to RNN training. In this paper, we adopt the boxcar function as our surrogate function. This choice has been proven effective in various contexts[9,11,67] and requires minimal computational resources as expressed by the derivative definition,

$$\frac{\partial s_t}{\partial u_t} = \begin{cases} 0.5 & \text{if} \quad |u_t - 1| \leq 0.5 \\ 0 & \text{otherwise}\,. \end{cases} \tag{10}$$

Overview of the auditory pathway in the brain

Sound waves are initially received by the outer ear and then transmitted as vibrations to the cochlea in the inner ear, where the basilar membrane allows for a representation of different frequencies along its length. Distinct sound frequencies induce localised membrane vibrations that activate adjacent inner hair cells. These specialised sensory cells, covering the entire basilar membrane, release neurotransmitters when activated, stimulating neighbouring auditory nerve fibers and initiating the production of action potentials. Tonotopy is maintained through the conversion of mechanical motion into electric signals as each inner hair cell, tuned to a specific frequency, only affects nearby auditory nerve fibers. The resulting spike trains then propagate through a multi-layered neural network, ultimately reaching cortical regions associated with higher-order cognitive functions such as speech recognition. Overall, the auditory system is organised hierarchically, with each level contributing to the progressively more sophisticated processing of auditory information.

Simulated speech recognition pipeline

Our objective is to design a speech recognition architecture that, while sufficiently plausible for meaningful comparisons with neuroscience observations, remains simple and efficient to ensure compatibility with modern deep learning techniques and achieve good ASR performance. We implement the overall waveform-to-phoneme pipeline illustrated in Fig. 1 inside the Speechbrain[68] framework. We provide a description of each of its components here below.

Feature extractor

80 Mel filterbank features are extracted with a 25 ms window and a shift of 2 ms, which down samples the 16 kHz input speech waveform to a 500 Hz spectrogram.

Auditory CNN

A single-layered two-dimensional convolution module is applied to the 80 extracted Mel features using 16 channels, a kernel size of (7, 7), a padding of (7, 0) and a stride of 1, producing $16 \cdot \big((80-7)+1\big) = 1184$ output signals with unchanged number of time steps. Layer normalisation, drop out on the channel dimension and a Leaky-ReLU activation are then applied. Each produced signal characterises the evolution over time of the spectral energy across a frequency band of 7 consecutive Mel bins.

Auditory nerve fibers

Each 500 Hz output signal from the auditory CNN constitutes the stimulus of a single auditory nerve fiber, which converts the real-valued signal into a spike train. These nerve fibers are modelled as a layer of LIF neurons without recurrent connections and using a single trainable parameter per neuron, $\tau_u \in [3, 25]$ ms, representing the time constant of the membrane potential decay.

Multi-layered SNN

The resulting spike trains are sent to a fully connected multi-layered SNN architecture with 512 neurons in each layer. The proportion of neurons with nonzero adaptation parameters is controlled in each layer so that only a fraction of the neurons are AdLIF and the rest are LIF. Similarly the proportion of nonzero feedforward and recurrent connections is controlled in each layer by applying fixed random binary masks to the weight matrices. Compared to a LIF neuron, an AdLIF neuron has three additional trainable parameters, $\tau_w \in [30, 350]$ ms, $a \in [-0.5, 5]$ and $b \in [0, 2]$, related to the adaptation variable coupled to the membrane potential.

Spikes to probabilities

The spike trains of the last layer are sent to a an average pooling module which down samples their time dimension to 25 Hz. These are then projected to 512 phoneme features using two FC layers with Leaky-ReLU activation. A third FC layer with Log-Softmax activation finally projects them to 40 log-probabilities representing 39 phoneme classes and a blank token as required by connectionist temporal classification (CTC).

Training and inference

The log-probabilities are sent to a CTC loss[69] so that the parameters of the complete architecture can be updated through back propagation. Additionally, some regularisation of the firing rate is used to prevent neurons from being silent or firing above the Nyquist frequency. At inference, CTC decoding is used to output the most likely phoneme sequence from the predicted log-probabilities, and the PER is computed to evaluate the model's performance.

Physiological plausibility and limitations

Cochlea and inner hair cells

While some of the complex biological processes involved in converting mechanical vibrations to electric neuron stimuli can be abstracted, we assume that the key feature to retain is the tonotopic encoding of sound information. A commonly used metric in neuroscience is the ratio of characteristic frequency to bandwidth, which defines how sharply tuned a neuron is around the frequency it is most responsive to. As detailed in Supplementary Table 4, from measured Q10 values in normal hearing humans reported in Devi *et al.*[70], we evaluate that a single auditory nerve fiber should receive inputs from 5-7 adjacent frequency bins when using 80 Mel filterbank features. The adoption of a Mel filterbank frontend can be justified by its widespread utilisation within deep learning ASR frameworks. Although we do not attempt to directly model cochlear and hair cell processing, we can provide a rough analogue in the form of Mel features passing through a trainable convolution module that yields plausible ranges of frequency sensitivity for our auditory nerve fibers.

Simuation time step

Modern ASR systems[1,4] typically use a frame period of $\Delta t = 10$ ms during feature extraction, which is then often sub-sampled to 40 ms using a CNN before entering the encoder-decoder architecture. In the brain, typical minimal inter-spike distances imposed by a neuron's absolute refractory period can vary from 0 to 5 ms[65]. We therefore assume that using a time step greater than 5 ms could result in dynamics that are less representative of biological phenomena. Although using a time step $\Delta t < 1$ ms may yield biologically more realistic simulations, we opt for time steps ranging from 1 to 5 ms to ensure computational efficiency. After the SNN, the spike trains of the last layer are down-sampled to 25 Hz via average pooling on the time dimension. This prevents an excessive number of time steps from entering the CTC loss, which could potentially hinder its decoding efficacy. We use $\Delta t = 5$ ms for most of the hyperparameter tuning to reduce training time, but favour $\Delta t = 2$ ms for the oscillation analysis so that the full gamma range of interest (30-150 Hz) remains below the Nyquist frequency at 250 Hz.

Neuron model

The LIF neuron model is an effective choice for modelling auditory nerve fibers as it accurately represents their primary function of encoding sensory inputs into spike trains. We avoid using SFA and recurrent connections, as they are not prevalent characteristics of nerve fibers. On the other hand, for the multi-layered SNN, the linear AdLIF neuron model with layer-wise recurrent connections stands out as a good compromise between accurately reproducing biological firing patterns and remaining computationally efficient[71].

Organisation in layers

Similarly to ANNs, our simulation incorporates a layered organisation, which facilitates the progressive extraction and representation of features from low-order to higher-order, without the need of concretely defining and distinguishing neuron populations. This fundamental architectural principle aligns with the hierarchical processing observed in biological brains and is justified by its compatibility with deep learning frameworks.

Layer-wise recurrence

While biological efferent pathways in the brain involve complex and widespread connections that span across layers and regions, modelling such intricate connectivity can introduce computational challenges and complexity, potentially hindering training and scalability. By restricting feedback connections to layer-wise recurrence, we simplify the network architecture and enhance compatibility with deep learning frameworks.

Excitatory and inhibitory

In the neuroscience field, neurons are commonly categorized into two types: excitatory neurons, which stimulate action potentials in postsynaptic neurons, and inhibitory neurons, which reduce the likelihood of spike production in postsynaptic neurons. In reality, neurons themselves are not always strictly excitatory or inhibitory, since a single neuron's axon branches can terminate on various target neurons, releasing either excitatory or inhibitory neurotransmitters. In ANNs, weight matrices are typically initialized with zero mean and a symmetric distribution, so that the initial number of excitatory and inhibitory connections is balanced. For the sake of compatibility with deep learning methods, we simply train the synaptic connections in all layers without constraining them to be excitatory or inhibitory and report what the proportion of excitatory-inhibitory connections converges to.

Learning rule

While stochastic gradient descent is biologically implausible due to its global and offline learning framework, it allows us to leverage parallelisable and fast computations to optimise larger-scale neural networks.

Decoding into phoneme sequences

Although lower PERs could be achieved with a more sophisticated decoder, our primary focus is on analysing the spiking layers within the encoder. For simplicity, we therefore opt for straightforward CTC decoding.

Hybrid ANN-SNN balance

The CNN module in the ASR frontend as well as the ANN module (average pooling and FC layers) converting spikes to probabilities are intentionally kept simple to give most of the processing and representational power to the central SNN on which focuses our neural oscillations analysis.

Speech processing tasks

The following datasets are used in our study.

- The TIMIT dataset[50] provides a comprehensive and widely utilised collection of phonetically balanced American English speech recordings from 630 speakers with detailed phonetic transcriptions and word alignments. It represents a standardised benchmark for evaluating ASR model performance. The training, validation and test sets contain 3696, 400 and 192 sentences respectively. Utterance durations vary between 0.9 to 7.8 seconds. Due to its compact size of approximately five hours of speech data, the TIMIT dataset is well-suited for investigating suitable model architectures and tuning hyperparameters. It is however considered small for ASR hence the use of Librispeech presented below.
- The Librispeech[56] corpus contains about 1,000 hours of English speech audiobook data read by over 2,400 speakers with utterance durations between 0.8 and 35 seconds. Given its significantly larger size, we only train a few models selected on TIMIT to confirm that our analysis holds when scaling to more data.
- The Google Speech Commands dataset[57] contains one-second audio recordings of 35 spoken commands such as "yes," "no," "stop," "go," "left," "right" "up". The training, validation and testing splits contain approximately 85k, 10k and 5k examples respectively. It is used to test whether similar CFCs arise when simply recognising single words instead of phoneme or subword sequences.

When evaluating an ASR model, the error rate signifies the proportion of incorrectly predicted words or phonemes. The count of successes in a binary outcome experiment, such as ASR testing, can be effectively modeled using a binomial distribution. In the context of trivial priors, the posterior distribution of the binomial distribution follows a beta distribution. By leveraging the equal-tailed 95% credible intervals derived from the posterior distribution, we establish error bars, yielding a range of $\pm 0.8\%$ for the reported PERs on the TIMIT dataset, about $\pm 0.2\%$ for the reported word error rates on LibriSpeech, and about $\pm 0.4\%$ for the reported accuracy on Google Speech Commands.

Analysis methods

Hyperparameter tuning

Before reporting results on the oscillation analysis, we investigate the optimal architecture by tuning some relevant hyperparameters. All experiments are run using a single NVIDIA GeForce RTX 3090 GPU. On top of assessing their respective impact on the error rate, we test if more physiologically plausible design choices correlate with better performance. Here is the list of the fixed parameters that we do not modify in our reported experiments:

- number of Mel bins: 80
- Mel window size: 25 ms
- auditory CNN kernel size (7, 7)
- auditory CNN stride: (1, 1)
- auditory CNN padding: (7, 0)
- average pooling size: 40 / Δt
- number of phoneme FC layers: 2
- number of phoneme features: 512
- dropout: 0.15
- activation: LeakyReLU

where for CNN attributes of the form (n_t, n_f), n_t and n_f correspond to time and feature dimensions respectively. The tunable parameters are the following,

- filter bank hop size controlling the SNN time step in ms: {1, 2, 5}
- number of auditory CNN channels (filters): {8, 16, 32, 64, 128}

- number of SNN layers: {1, 2, 3, 4, 5, 6, 7}
- neurons per SNN layer: {64, 128, 256, 512, 768, 1024, 1536, 2048}
- proportion of neurons with SFA: [0, 1]
- feedforward connectivity: [0, 1]
- recurrent connectivity: [0, 1]

While increasing the number of neurons per layer primarily impacts memory requirements, additional layers mostly extend training time.

Population signal

In the neuroscience field, EEG stands out as a widely employed and versatile method for studying brain activity. By placing electrodes on the scalp, this non-invasive technique measures the aggregate electrical activity resulting from the synchronised firing of neurons within a specific brain region. An EEG signal therefore reflects the summation of postsynaptic potentials from a large number of neurons operating in synchrony. The typical sampling rate for EEG data is commonly in the range of 250 to 1000 Hz which matches our desired simulation time steps. With our SNN, we do not have EEG signals but directly the individual spike trains of all neurons in the architecture. In order to perform similar population-level analyses, we simply sum the binary spike trains $S^l \in \{0,1\}^{T \times N^l}$ emitted by all neurons in a specific layer l. The resulting population signals are then normalised with a mean of 0 and a standard deviation of 1 before performing the PAC analysis described below.

Phase-amplitude coupling

Using finite impulse response band-pass filters, the obtained population signals are decomposed into different frequency ranges. We study CFC in the form of PAC both within a single population and across layers. This technique assesses whether a relationship exists between the phase of a low frequency signal and the envelope (amplitude) of a high frequency signal. As recommended by Hulsemann *et al.*[72], we implement both the modulation index[73] and mean vector length[38] metrics to quantify the observed amount of PAC. For each measure type, the observed coupling value is compared to a distribution of 10,000 surrogates to assess the significance. A surrogate coupling is computed between the original phase time series and a permuted amplitude time series, constructed by cutting at a random data point and reversing the order of both parts. A p-value can then be obtained by fitting a Gaussian function on the distribution of surrogate coupling values and calculating the area under the curve for values greater than the observed coupling value. As pointed out by Jones[74], it is important to note that observed oscillations can exhibit complexities such as non-sinusoidal features and brief transient events on single trials. Such nuances become aggregated when averaging signals, leading to the widely observed continuous rhythms. We therefore perform all analysis on single utterances. For intra-layer interactions, a single population signal is used to extract both the low-frequency oscillation phase and the high-frequency oscillation amplitude. In a three-layered architecture, they include nerve-nerve, first layer-first layer, second layer-second layer, and third layer-third layer couplings. For inter-layer interactions, we consider couplings between the low-frequency oscillation phase in a given layer and the high-frequency oscillation amplitude in all subsequent layers. These consist of nerve-first layer, nerve-second layer, nerve-third layer, first layer-second layer, first layer-third layer, second layer-third layer couplings. For all aforementioned intra- and inter-layer combinations, we use delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz) and beta (13-30 Hz) ranges as low-frequency modulating bands, and low-gamma (30-80 Hz) and high-gamma (80-150 Hz) ranges as high-frequency modulated bands. For a given model, we iterate through the 64 longest utterances in the TIMIT test set. For each utterance, we consider the 10 aforementioned intra- and inter-layer relations, as well as the 8 possible combinations of low-frequency to high-frequency bands. We conduct PAC testing on each of the 5,120 resulting coupling scenarios, and only consider a coupling to be significant when both modulation index and mean vector length metrics yield p-values below 0.05.

Data availability

All datasets used to conduct our study are publicly available. The TIMIT dataset, along with relevant access information, can be found on the Linguistic Data Consortium website at https://catalog.ldc.upenn.edu/LDC93S1. The Librispeech and Google Speech Commands datasets are directly available at https://www.openslr.org/12. and https://www.tensorflow.org/datasets/catalog/speech_commands respectively.

Code availability

To facilitate the advancement of spiking neural networks, we have made our code open source. The code repository will be accessible via a link in place of this text upon publication.

References

1. Gulati, A. *et al.* Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech*, 5036–5040, DOI: 10.21437/Interspeech.2020-3015 (2020).

2. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. neural information processing systems* **33**, 12449–12460 (2020).

3. Li, B. *et al.* Scaling end-to-end models for large-scale multilingual ASR. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1011–1018 (IEEE, 2021).

4. Radford, A. *et al.* Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518 (PMLR, 2023).

5. Brodbeck, C., Hannagan, T. & Magnuson, J. S. Recurrent neural networks as neuro-computational models of human speech recognition. *bioRxiv* 2024–02, DOI: 10.1101/2024.02.20.580731 (2024).

6. Millet, J. *et al.* Toward a realistic model of speech processing in the brain with self-supervised learning. *Adv. Neural Inf. Process. Syst.* **35**, 33428–33443 (2022).

7. Millet, J. & King, J.-R. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech (2021). 2103.01032.

8. Magnuson, J. S. *et al.* Earshot: A minimal neural network model of incremental human speech recognition. *Cogn. science* **44**, e12823, DOI: 10.1111/cogs.12823 (2020).

9. Bittar, A. & Garner, P. N. A surrogate gradient spiking baseline for speech command recognition. *Front. Neurosci.* **16**, DOI: 10.3389/fnins.2022.865897 (2022).

10. Neftci, E. O., Mostafa, H. & Zenke, F. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Process. Mag.* **36**, 51–63, DOI: 10.1109/MSP.2019.2931595 (2019).

11. Bittar, A. & Garner, P. N. Surrogate gradient spiking neural networks as encoders for large vocabulary continuous speech recognition (2022). 2212.01187.

12. Buzsaki, G. *Rhythms of the Brain* (Oxford university press, 2006).

13. Fries, P., Reynolds, J. H., Rorie, A. E. & Desimone, R. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291**, 1560–1563, DOI: 10.1126/science.1055465 (2001).

14. Jensen, O. & Colgin, L. L. Cross-frequency coupling between neuronal oscillations. *Trends cognitive sciences* **11**, 267–269, DOI: 10.1016/j.tics.2007.05.003 (2007).

15. Womelsdorf, T. & Fries, P. The role of neuronal synchronization in selective attention. *Curr. opinion neurobiology* **17**, 154–160, DOI: 10.1016/j.conb.2007.02.002 (2007).

16. Vinck, M., Womelsdorf, T., Buffalo, E. A., Desimone, R. & Fries, P. Attentional modulation of cell-class-specific gamma-band synchronization in awake monkey area v4. *Neuron* **80**, 1077–1089, DOI: 10.1016/j.neuron.2013.08.019 (2013).

17. Kucewicz, M. T. *et al.* Dissecting gamma frequency activity during human memory processing. *Brain* **140**, 1337–1350, DOI: 10.1093/brain/awx043 (2017).

18. Başar, E., Başar-Eroğlu, C., Karakaş, S. & Schürmann, M. Brain oscillations in perception and memory. *Int. journal psychophysiology* **35**, 95–124, DOI: 10.1016/S0167-8760(99)00047-1 (2000).

19. Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S. & Woldorff, M. G. Good times for multisensory integration: effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia* **45**, 561–571, DOI: 10.1016/j.neuropsychologia.2006.01.013 (2007).

20. MacKay, W. A. Synchronized neuronal oscillations and their role in motor processes. *Trends cognitive sciences* **1**, 176–183, DOI: 10.1016/S1364-6613(97)01059-0 (1997).

21. Ramos-Murguialday, A. & Birbaumer, N. Brain oscillatory signatures of motor tasks. *J. neurophysiology* **113**, 3663–3682, DOI: 10.1152/jn.00467.2013 (2015).

22. Jirsa, V. & Müller, V. Cross-frequency coupling in real and virtual brain networks. *Front. computational neuroscience* **7**, 78, DOI: 10.3389/fncom.2013.00078 (2013).

23. Abubaker, M., Al Qasem, W. & Kvašňák, E. Working memory and cross-frequency coupling of neuronal oscillations. *Front. Psychol.* **12**, 756661, DOI: 10.3389/fpsyg.2021.756661 (2021).

24. Tort, A. B., Komorowski, R. W., Manns, J. R., Kopell, N. J. & Eichenbaum, H. Theta–gamma coupling increases during the learning of item–context associations. *Proc. Natl. Acad. Sci.* **106**, 20942–20947, DOI: 10.1073/pnas.0911331106 (2009).

25. Axmacher, N. *et al.* Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proc. Natl. Acad. Sci.* **107**, 3228–3233, DOI: 10.1073/pnas.0911531107 (2010).

26. Buzsáki, G. & Moser, E. I. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat. neuroscience* **16**, 130–138, DOI: 10.1038/nn.3304 (2013).

27. Backus, A. R., Schoffelen, J.-M., Szebényi, S., Hanslmayr, S. & Doeller, C. F. Hippocampal-prefrontal theta oscillations support memory integration. *Curr. Biol.* **26**, 450–457, DOI: 10.1016/j.cub.2015.12.048 (2016).

28. Hummos, A. & Nair, S. S. An integrative model of the intrinsic hippocampal theta rhythm. *PloS one* **12**, e0182648, DOI: 10.1371/journal.pone.0182648 (2017).

29. Reddy, L. *et al.* Theta-phase dependent neuronal coding during sequence learning in human single neurons. *Nat. communications* **12**, 4839, DOI: 10.1038/s41467-021-25150-0 (2021).

30. Colgin, L. L. Mechanisms and functions of theta rhythms. *Annu. review neuroscience* **36**, 295–312, DOI: 10.1146/annurev-neuro-062012-170330 (2013).

31. Itskov, V., Pastalkova, E., Mizuseki, K., Buzsaki, G. & Harris, K. D. Theta-mediated dynamics of spatial information in hippocampus. *J. Neurosci.* **28**, 5959–5964, DOI: 10.1523/JNEUROSCI.5262-07.2008 (2008).

32. Mizuseki, K., Sirota, A., Pastalkova, E. & Buzsáki, G. Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. *Neuron* **64**, 267–280, DOI: 10.1016/j.neuron.2009.08.037 (2009).

33. Foxe, J. J. & Snyder, A. C. The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Front. psychology* **2**, 10747, DOI: 10.3389/fpsyg.2011.00154 (2011).

34. Banerjee, S., Snyder, A. C., Molholm, S. & Foxe, J. J. Oscillatory alpha-band mechanisms and the deployment of spatial attention to anticipated auditory and visual target locations: supramodal or sensory-specific control mechanisms? *J. Neurosci.* **31**, 9923–9932, DOI: 10.1523/JNEUROSCI.4660-10.2011 (2011).

35. Jensen, O. & Mazaheri, A. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front. human neuroscience* **4**, 186, DOI: 10.3389/fnhum.2010.00186 (2010).

36. Klimesch, W. Alpha-band oscillations, attention, and controlled access to stored information. *Trends cognitive sciences* **16**, 606–617, DOI: 10.1016/j.tics.2012.10.007 (2012).

37. Engel, A. K., Fries, P. & Singer, W. Dynamic predictions: oscillations and synchrony in top–down processing. *Nat. Rev. Neurosci.* **2**, 704–716, DOI: 10.1038/35094565 (2001).

38. Canolty, R. T. *et al.* High gamma power is phase-locked to theta oscillations in human neocortex. *science* **313**, 1626–1628, DOI: 10.1126/science.1128115 (2006).

39. Ghitza, O. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front. psychology* **2**, 130, DOI: 10.3389/fpsyg.2011.00130 (2011).

40. Giraud, A.-L. & Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. neuroscience* **15**, 511–517, DOI: 10.1038/nn.3063 (2012).

41. Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B. & Giraud, A.-L. Speech encoding by coupled cortical theta and gamma oscillations. *elife* **4**, e06213, DOI: 10.7554/eLife.06213 (2015).

42. Attaheri, A. *et al.* Delta-and theta-band cortical tracking and phase-amplitude coupling to sung speech by infants. *NeuroImage* **247**, 118698, DOI: 10.1016/j.neuroimage.2021.118698 (2022).

43. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010, DOI: 10.1126/science.1245994 (2014).

44. Bonhage, C. E., Meyer, L., Gruber, T., Friederici, A. D. & Mueller, J. L. Oscillatory EEG dynamics underlying automatic chunking during sentence processing. *NeuroImage* **152**, 647–657, DOI: 10.1016/j.neuroimage.2017.03.018 (2017).

45. Strauß, A., Wöstmann, M. & Obleser, J. Cortical alpha oscillations as a tool for auditory selective inhibition. *Front. human neuroscience* **8**, 350, DOI: 10.3389/fnhum.2014.00350 (2014).

46. Strauß, A., Kotz, S. A., Scharinger, M. & Obleser, J. Alpha and theta brain oscillations index dissociable processes in spoken word recognition. *Neuroimage* **97**, 387–395, DOI: 10.1016/j.neuroimage.2014.04.005 (2014).

47. Wöstmann, M., Lim, S.-J. & Obleser, J. The human neural alpha response to speech is a proxy of attentional control. *Cereb. cortex* **27**, 3307–3317, DOI: 10.1093/cercor/bhx074 (2017).

48. Obleser, J. & Weisz, N. Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cereb. cortex* **22**, 2466–2477, DOI: 10.1093/cercor/bhr325 (2012).

49. Hovsepyan, S., Olasagasti, I. & Giraud, A.-L. Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nat. communications* **11**, 3117, DOI: 10.1038/s41467-020-16956-5 (2020).

50. Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G. & Pallett, D. S. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n* **93**, 27403, DOI: 10.35111/17gk-bn40 (1993).

51. Perez-Nieves, N., Leung, V. C., Dragotti, P. L. & Goodman, D. F. Neural heterogeneity promotes robust learning. *Nat. Commun.* **12**, 5791, DOI: 10.1038/s41467-021-26022-3 (2021).

52. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* **5**, 157–166, DOI: 10.1109/72.279181 (1994).

53. Crook, S. M., Ermentrout, G. B. & Bower, J. M. Spike frequency adaptation affects the synchronization properties of networks of cortical oscillators. *Neural Comput.* **10**, 837–854, DOI: 10.1162/089976698300017511 (1998).

54. Augustin, M., Ladenbauer, J. & Obermayer, K. How adaptation shapes spike rate oscillations in recurrent neuronal networks. *Front. computational neuroscience* **7**, 9, DOI: 10.3389/fncom.2013.00009 (2013).

55. Reddy, C. K. *et al.* A scalable noisy speech dataset and online subjective test framework. *Proc. Interspeech 2019* 1816–1820, DOI: 10.21437/Interspeech.2019-3087 (2019).

56. Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210 (IEEE, 2015).

57. Warden, P. Speech commands: A dataset for limited-vocabulary speech recognition (2018). 1804.03209.

58. Hammouamri, I., Khalfaoui-Hassani, I. & Masquelier, T. Learning delays in spiking neural networks using dilated convolutions with learnable spacings (2023). 2306.17670.

59. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The J. physiology* **117**, 500, DOI: 10.1113/jphysiol.1952.sp004764 (1952).

60. FitzHugh, R. Impulses and physiological states in theoretical models of nerve membrane. *Biophys. journal* **1**, 445, DOI: 10.1016/s0006-3495(61)86902-6 (1961).

61. Morris, C. & Lecar, H. Voltage oscillations in the barnacle giant muscle fiber. *Biophys. journal* **35**, 193–213, DOI: 10.1016/S0006-3495(81)84782-0 (1981).

62. Izhikevich, E. M. Simple model of spiking neurons. *IEEE Transactions on neural networks* **14**, 1569–1572, DOI: 10.1109/TNN.2003.820440 (2003).

63. Brette, R. & Gerstner, W. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J. neurophysiology* **94**, 3637–3642, DOI: 10.1152/jn.00686.2005 (2005).

64. Badel, L. *et al.* Extracting non-linear integrate-and-fire models from experimental data using dynamic i–v curves. *Biol. cybernetics* **99**, 361–370, DOI: 10.1007/s00422-008-0259-4 (2008).

65. Gerstner, W. & Kistler, W. M. *Spiking neuron models: Single neurons, populations, plasticity* (Cambridge university press, 2002).

66. Paszke, A. *et al.* Automatic differentiation in pytorch. *NIPS Work.* (2017).

67. Kaiser, J., Mostafa, H. & Neftci, E. Synaptic plasticity dynamics for deep continuous local learning (DECOLLE). *Front. Neurosci.* **14**, 424, DOI: 10.3389/fnins.2020.00424 (2020).

68. Ravanelli, M. *et al.* SpeechBrain: A general-purpose speech toolkit (2021). 2106.04624.

69. Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376, DOI: 10.1145/1143844.1143891 (2006).

70. Devi, N., Sreeraj, K., Anuroopa, S., Ankitha, S. & Namitha, V. Q10 and tip frequencies in individuals with normal-hearing sensitivity and sensorineural hearing loss. *Indian J. Otol.* **28**, 126–129, DOI: 10.4103/indianjotol.indianjotol_5_22 (2022).

71. Ganguly, C. *et al.* Spike frequency adaptation: bridging neural models and neuromorphic applications. *Commun. Eng.* **3**, 22, DOI: 10.1038/s44172-024-00165-9 (2024).

72. Hülsemann, M. J., Naumann, E. & Rasch, B. Quantification of phase-amplitude coupling in neuronal oscillations: comparison of phase-locking value, mean vector length, modulation index, and generalized-linear-modeling-cross-frequency-coupling. *Front. neuroscience* **13**, 573, DOI: 10.3389/fnins.2019.00573 (2019).

73. Tort, A. B. *et al.* Dynamic cross-frequency couplings of local field potential oscillations in rat striatum and hippocampus during performance of a T-maze task. *Proc. Natl. Acad. Sci.* **105**, 20517–20522, DOI: 10.1073/pnas.0810524105 (2008).

74. Jones, S. R. When brain rhythms aren't 'rhythmic': implication for their mechanisms and meaning. *Curr. opinion neurobiology* **40**, 72–80, DOI: 10.1016/j.conb.2016.06.010 (2016).

Acknowledgements

This project received funding under NAST: Neural Architectures for Speech Technology, Swiss National Science Foundation grant 185010 (https://data.snf.ch/grants/grant/185010).

Author contributions statement

AB designed and implemented the architecture, performed training and analysis experiments, did most of the literature search, and wrote the bulk of the manuscript. PG secured funding, supervised the experimental work, and assisted with literature and writing. All authors contributed to the article and approved the submitted version.

Corresponding author

Correspondence and requests for materials should be addressed to Alexandre Bittar at abittar@idiap.ch.

Competing interests

The authors declare no competing interests.

Supplementary Material

Appendix

Forward-Euler first-order exponential integrator method

Consider the following differential equation,

$$\tau \dot{y} = -y + \mathscr{N}(y) + \tau \gamma s(t)\,, \tag{1}$$

where $\mathscr{N}(y)$ is a nonlinear function and $s(t)$ is a spike train defined as

$$s(t) = \sum_f \delta(t - t^f)\,. \tag{2}$$

Multiplying both sides by $\frac{1}{\tau}\exp(\frac{t}{\tau})$ and then integrating over $[t_n, t_{n+1}]$, where $t_n = n\Delta t$, yields

$$\int_{t_n}^{t_{n+1}} \left(\dot{y}\exp\frac{t}{\tau} + y\frac{1}{\tau}\exp\frac{t}{\tau}\right)dt = \frac{1}{\tau}\int_{t_n}^{t_{n+1}} \mathscr{N}(y)\exp\frac{t}{\tau}\,dt + \gamma\sum_f \int_{t_n}^{t_{n+1}} \exp\frac{t}{\tau}\,\delta(t-t^f)\,. \tag{3}$$

The left hand side has an exact solution

$$y(t)\exp\frac{t}{\tau}\bigg|_{t=t_n}^{t_{n+1}} = \left(y_{n+1}\exp\frac{\Delta t}{\tau} - y_n\right)\exp\frac{t_n}{\tau}\,. \tag{4}$$

For the first term of the right hand side, the nonlinearity $\mathscr{N}(y)$ can be approximated as constant over $[t_n, t_{n+1}]$ for sufficiently small Δt, so that $\mathscr{N}(y) \approx \mathscr{N}(y_n)$ and we can solve it as

$$\frac{1}{\tau}\int_{t_n}^{t_{n+1}} \mathscr{N}(y)\exp\frac{t}{\tau}\,dt \approx \mathscr{N}(y_n)\exp\frac{t}{\tau}\bigg|_{t=t_n}^{t_{n+1}} = \mathscr{N}(y_n)\exp\frac{t_n}{\tau}\left(\exp\frac{\Delta t}{\tau} - 1\right). \tag{5}$$

Finally for the last term, the width Δt of the interval $[t_n, t_{n+1}]$ can be set sufficiently small to include at most a single spike. The exact firing time $t^f \in [t_n, t_{n+1}]$ can then be discretised as $t^f = t_n$ so that $s_n = \sum_f \delta(t_n - t^f)$ and

$$\gamma\sum_f \exp\frac{t^f}{\tau}\bigg|_{t^f\in[t_n,t_{n+1}]} = \gamma\exp\frac{t^n}{\tau}\,s_n \tag{6}$$

Putting everything together, we get the following update equation for y in discrete time,

$$y_{n+1} = \exp\frac{-\Delta t}{\tau}\left(y_n + \gamma s_n\right) + \left(1 - \exp\frac{-\Delta t}{\tau}\right)\mathscr{N}(y_n)\,. \tag{7}$$

Eigenvalues of AdLIF free equations

The free equations of the AdLIF neuron model are obtained by considering Eqs. (3) and (4) in the special case where there is no input, $I(t) = 0$, and no emitted spikes, $s(t) = 0$. They can be rewritten in matrix form as,

$$\frac{d}{dt}\begin{bmatrix} u \\ w \end{bmatrix} = \begin{bmatrix} -1/\tau_u & -1/\tau_u \\ a/\tau_w & -1/\tau_w \end{bmatrix}\begin{bmatrix} u \\ w \end{bmatrix} = A\begin{bmatrix} u \\ w \end{bmatrix}. \tag{8}$$

The eigenvalues can be found by setting the determinant of $A - \lambda\mathbb{I}$ to zero,

$$\begin{vmatrix} -1/\tau_u - \lambda & -1/\tau_u \\ a/\tau_w & -1/\tau_w - \lambda \end{vmatrix} = 0\,, \tag{9}$$

yielding the characteristic polynomial,

$$\lambda^2 + \lambda\left(\frac{1}{\tau_u} + \frac{1}{\tau_w}\right) + \frac{1+a}{\tau_u\tau_w} = 0\,, \tag{10}$$

whose roots correspond to the two eigenvalues of the system,

$$\lambda_{1,2} = -\frac{1}{2}\left(\frac{1}{\tau_u} + \frac{1}{\tau_w}\right) \pm \frac{1}{2}\sqrt{\left(\frac{1}{\tau_u} + \frac{1}{\tau_w}\right)^2 - \frac{4(1+a)}{\tau_u\tau_w}}\,. \tag{11}$$

In order to prevent the occurrence of exponentially growing solutions and ensure stability, both eigenvalues need to have a strictly negative real part, which can be realised by imposing a lower bound $a > -1$ on the coupling strength. Moreover, allowing eigenvalues to have a nonzero imaginary part introduces the potential for oscillatory modes that may amplify perturbations. This could cause some challenges in terms of numerical stability, convergence and interpretability, especially in the context of deep neural networks trained with gradient descent. We therefore impose an additional upper bound on the values of a leading to the overall stability condition,

$$-1 < a \leq \frac{(\tau_w - \tau_u)^2}{4\tau_u\tau_w}\,. \tag{12}$$

Kernel formulation of a spiking neuron

Using the SRM formulation, the membrane potential $u(t)$ is described as,

$$u(t) = \int_0^\infty \kappa(s)\, I(t-s) ds + \int_0^\infty \eta(s)\, S(t-s) ds\,, \tag{13}$$

where the two kernels $\kappa(s)$ and $\eta(s)$ describe the response to an input pulse and the response to an afterspike reset pulse respectively. It can be shown that the differential equations of a linear AdLIF neuron has an equivalent kernel formulation with,

$$\kappa(s) = \left(\beta_1\, e^{\lambda_1 s} + \beta_2\, e^{\lambda_2 s}\right)\Theta(s) \tag{14a}$$

$$\eta(s) = \left(\gamma_1\, e^{\lambda_1 s} + \gamma_2\, e^{\lambda_2 s}\right)\Theta(s)\,, \tag{14b}$$

where λ_1, λ_2 are the eigenvalues of the system given in Supplementary Eq. (11) and $\Theta(s)$ is the Heaviside step function. The coefficients β_1, β_2 of the input kernel are such that the membrane potential increases by $\Delta u = 1$, without any effect on the recovery current, i.e., $\Delta w = 0$,

$$\beta_1 = \frac{\tau_u\lambda_2 + 1}{\tau_u(\lambda_2 - \lambda_1)} \quad \text{and} \quad \beta_2 = 1 - \beta_1\,. \tag{15}$$

The coefficients γ_1, γ_2 on the afterspike reset kernel are such that the membrane potential decreases by $\Delta u = \vartheta - u_r$ and the recovery current jumps by an amount $\Delta w = b$,

$$\gamma_1 = \frac{b - (\vartheta - u_r)(\tau_u\lambda_2 + 1)}{\tau_u(\lambda_2 - \lambda_1)} \quad \text{and} \quad \gamma_2 = -(\vartheta - u_r) - \gamma_1\,. \tag{16}$$

Supplementary Tables

Table 1. Hyperparameter tuning for the simulation time step on the TIMIT dataset. PERs are reported after 50 epochs using 16 CNN channels, 3 layers of 512 neurons each, 50% of AdLIF neurons, 100% feedforward and 50% recurrent connectivity.

Time step [ms]	Epoch duration [min]	Test PER [%]	Validation PER [%]
5	21	20.5	18.2
2	53	20.4	18.7
1	156	20.6	18.2

Table 2. Hyperparameter tuning for the number of CNN channels on the TIMIT dataset. PERs are reported after 50 epochs using a 5 ms time step, 3 layers, 512 neurons per layer, 50% of AdLIF neurons, 100% feedforward and 50% recurrent connectivity. Bold values indicate the lowest achieved PERs.

CNN channels	Nerve fibers	Parameters in complete encoder	Test PER [%]	Validation PER [%]
8	592	1.8M	20.9	18.9
16	1,184	2.1M	20.5	18.2
32	2,368	2.7M	20.2	18.4
64	4,736	3.9M	**19.8**	**18.0**
128	9,472	6.4M	21.3	18.9

Table 3. Hyperparameter tuning for the feedforward connectivity on the TIMIT dataset. PERs are reported after 50 epochs using a 5 ms time step, 16 CNN channels, 3 layers of 512 neurons each, 50% of AdLIF neurons and 50% recurrent connectivity. Bold values indicate the lowest achieved PERs.

Feedforward connectivity	Number of parameters	Test PER [%]	Validation PER [%]
0.2	629k / 1.2M	22.0	19.6
0.5	968k / 1.5M	21.6	19.7
0.8	1.3M / 1.8M	20.7	18.8
1.0	1.5M / 2.1M	**20.5**	**18.2**

Table 4. Sensitivity to frequency of auditory nerve fibers. The second column gives the average ranges measured by Devi *et al.* and the third the surrounding Mel scale centers when using 80 filters.

Characteristic Frequency [Hz]	Sensitivity range [Hz]	Nearby Mel bin centers [Hz]	Overlapping Number of bins
500	416-583	416, 452, 488, 525, 564, 604	5-6
1000	881-1119	872, 921, 973, 1026, 1080, 1136	5-6
2000	1770-2230	1729, 1806, 1886, 1967, 2052, 2139, 2228	6-7
4000	3566-4434	3553, 3688, 3827, 3970, 4117, 4270, 4427	7
6000	5501-6499	5479, 5674, 5875, 6083, 6297, 6519	5-6

Supplementary Figures

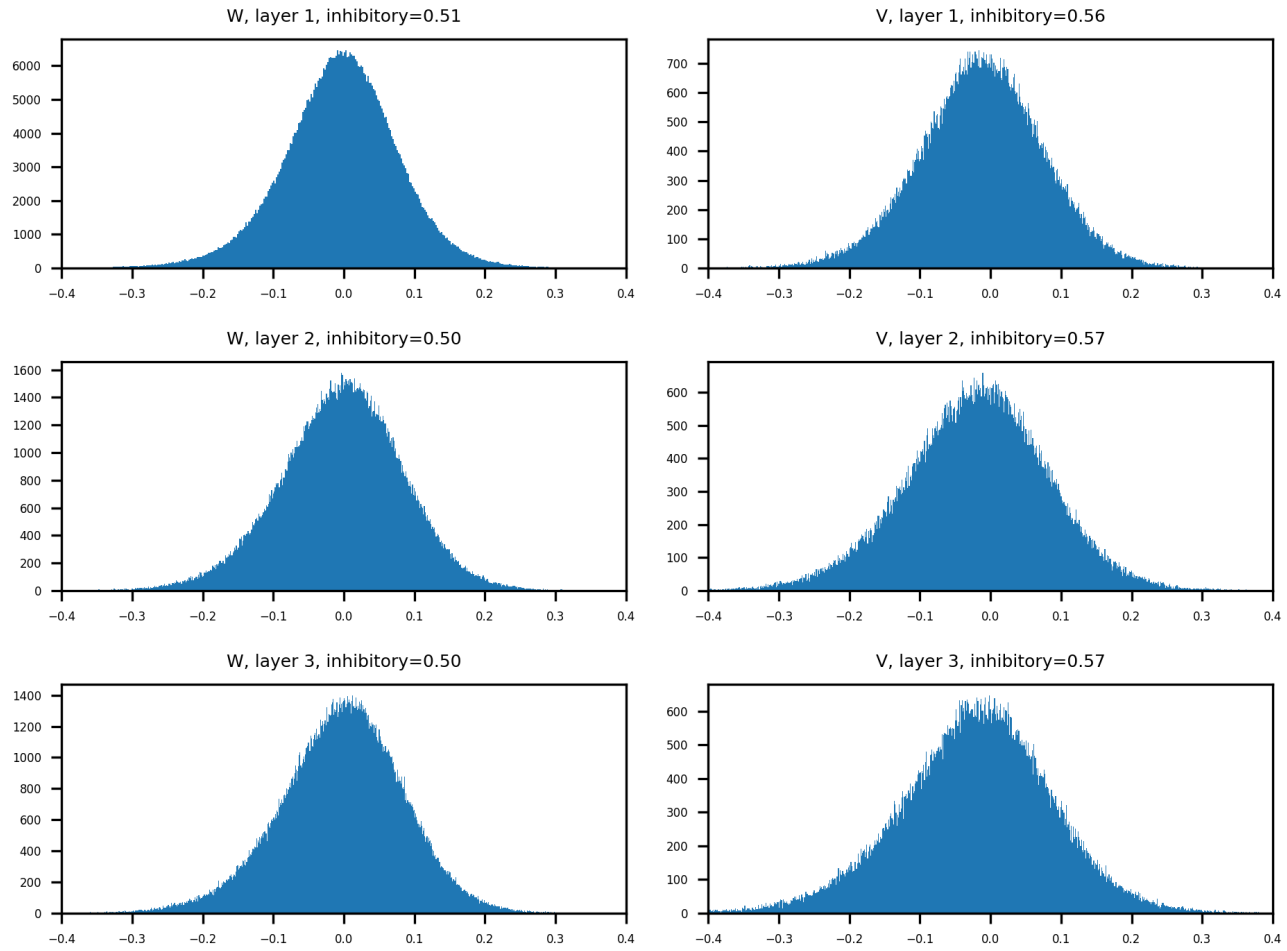

Figure 1. Distribution of feedforward and recurrent weight values across layers. The model uses a 2 ms time step, 16 CNN channels, 3 layers of size 512, 50% AdLIF neurons, 100% feedforward and 50% recurrent connectivity. After training, while feedforward connections do not have a dominant tendency, about 60% of the nonzero recurrent connections are inhibitory (negative valued).