Neural rhythms as priors of speech computations

Nand Chandravadia[†1,2] and Nabil Imam[‡1,2]

[1]School of Computational Science and Engineering, Georgia Institute of Technology
[2]Machine Learning Center (ML@GT), Georgia Institute of Technology

Abstract

The transformation of continuous speech into discrete linguistic representations forms the basis of speech recognition. Natural speech encodes cues at distinct timescales. Phonetic features have modulation frequencies of 30-50 Hz, syllables and words around 4-7 Hz, and phrases 1-2 Hz. Strikingly, these frequencies mirror frequencies of endogenous network rhythms of the brain and synaptic time constants of the underlying neural circuits. Here, we suggest that endogenous brain rhythms serve as priors for speech recognition, encoding knowledge of speech structure in the dynamics of network computations. In a network of coupled oscillators, we find that speech is readily identified when characteristic frequencies of the oscillators match frequencies of circuit rhythms in the brain. When signal and circuit rhythms are mismatched, speech identification is impaired. Compared to a baseline recurrent neural network without intrinsic oscillations, the coupled oscillatory network has significantly higher performance in speech recognition across languages, but not in the recognition of signals that lack speech-like structure, such as urban sounds. Our results suggest a central computational role of brain rhythms in speech processing.

Introduction

Parsing a continuous stream of speech within mixtures of unrelated sounds relies on coordinated neural processes in the brain that track temporal fluctuations of the speech envelope. The auditory cortex displays characteristic patterns of neural activity, notably low-frequency neural oscillations ("rhythms") in the delta (δ) and theta (θ) bands, and heightened sensitivity to stimulus modulations in this range [1–3]. Interestingly, temporal modulations of speech envelopes exhibit rhythms in a similar low frequency range, with envelope spectra typically peaking at around 4 Hz [4–6]. When the envelope is artificially compressed, speech comprehension significantly degrades, but recovers when silent gaps are inserted to restore the natural temporal structure of the signal [7]. These observations suggest that the natural frequencies of auditory neural circuits provide a temporal scaffold upon which speech computations take place.

Temporally structured neural activity patterns in the brain tile a range of roughly three orders of magnitude from ultra-slow (< 0.01 Hz) to ultra-fast (~ 600 Hz) oscillations [8]. Within this broad range are five functional oscillation classes – delta (δ),

[†]nchandravadia3@gatech.edu
[‡]nimam6@gatech.edu

theta (θ), alpha (α), beta (β), and gamma (γ) – whose interactions are presumed to subserve various functions, including speech and language processing. It has been proposed that networks of inhibitory and excitatory neurons underlying these oscillations give rise to larger computational units that function as coupled harmonic oscillators [4, 9–11]. These networks of oscillators exhibit periodic ("wavelike") behavior characterized by frequency, amplitude, and phase, with coupling strengths dynamically adjusted in response to external stimuli. Coupling between oscillators operating at different timescales mirror hierarchical temporal information integration in the brain.

Here, we use model networks of globally coupled oscillators [11, 12] trained with tools from machine learning to investigate endogenous network rhythms of the brain and their relationship to the natural rhythms of input speech signals. The dynamics of each network is defined by the characteristic frequency and damping coefficient of individual oscillators, as well as the their coupling strengths which are learned from speech data. These intrinsic and learned network parameters determine network trajectories over time, which are decoded to categorize speech utterances. We demonstrate that: (1) encoding *a priori* low-frequency brain-like rhythms in the oscillators significantly improve speech recognition performance compared to conventional recurrent neural networks, (2) these computational benefits are consistent across three spoken languages: English, Arabic, and Bengali, (3) network rhythms that align best with the temporal modulations of the speech envelope provide the greatest performance gains across all speech tasks, and (4) these benefits are absent in a non-speech task involving the recognition of urban sounds that lack the temporal structure of speech. These results suggest that low-frequency brain rhythms serve as a computational scaffold for broadband speech processing, where coupled oscillations act through the superposition and interference patterns of waves to efficiently extract structured temporal patterns in speech.

Results

Rhythms of Speech

The temporal modulations of the speech envelope have a peak energy concentration between 1-5 Hz [4, 13]. These modulations reflect the natural phrasal and syllabic rhythms of spoken language (Fig. 1A). The envelope, which represents the slow amplitude fluctuations of the speech signal over time, encodes the overall intensity pattern of speech and reflects essential rhythmic features of spoken language such as syllabic stress and phrasal timing. The hierarchical organization of language is captured in these low-frequency modulations, with slower components conveying phrase-level structure and faster ones conveying syllabic transitions. These enable the segmentation and interpretation of continuous speech.

Unlike the structured and hierarchical rhythms of speech, urban sounds encompass a wide variety of auditory patterns, including irregular noises like car horns or construction activity (Fig. 1B). These sounds lack the temporal regularity and predictable modulations characteristic of speech, instead featuring abrupt onsets, wide spectral distributions, and overlapping frequencies. For instance, while the speech envelope predominantly occupies low-frequency bands between 1-5 Hz, urban sounds span a broader range of frequencies with erratic amplitude fluctuations that lack hierarchical organization. The envelope of

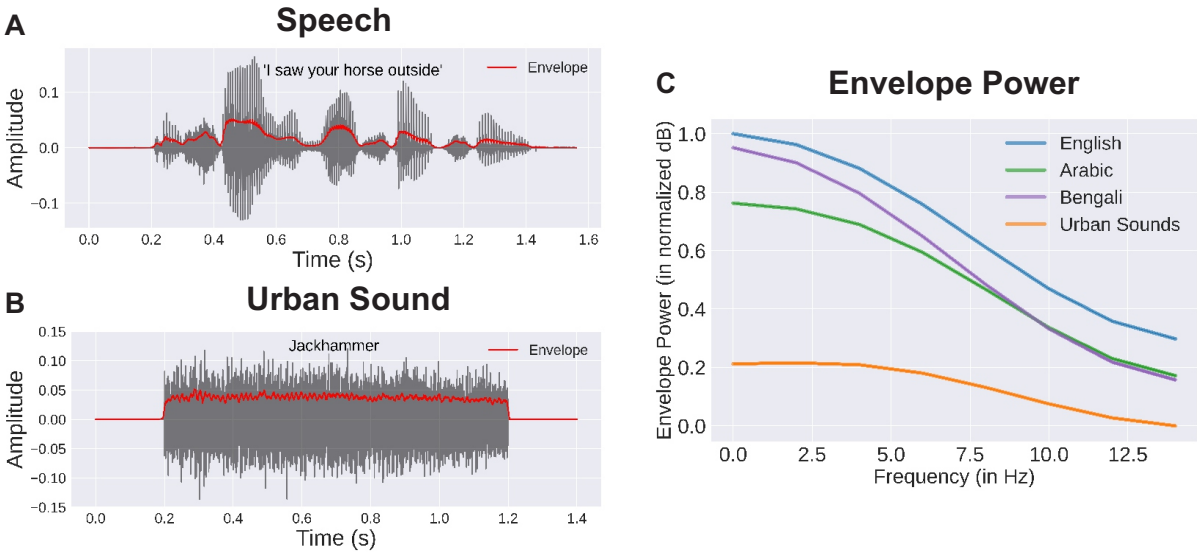

Figure 1: Dynamics of speech and non-speech stimuli. **(A)** Raw audio of the spoken sentence "I saw your horse outside" (grey) with the corresponding speech envelope (red). **(B)** Raw audio of an urban sound ('Jackhammer') (grey) and its envelope (red). **(C)** Average power (in normalized dB, $[0, 1]$) of the envelopes in three spoken language datasets – English, Arabic, and Bengali – and for a dataset of urban sounds.

urban sounds reflects these irregularities, with peaks and valleys occurring sporadically rather than in systematic rhythmic patterns.

To assess differences in the modulation spectra of speech and urban sounds, we analyze the envelope power across three spoken languages – English, Arabic, and Bengali – and compare it to that of urban sounds encountered in daily life (Fig. 1C). Across languages, we observe a distinctive pattern: the speech envelope peaks between the 1–5 Hz frequency range, reflecting the regular rate at which speakers of different languages convey linguistic information (Fig. 1C, English, Arabic, Bengali). In contrast, urban sounds lack a putative rhythmic pattern, exhibiting no clear peak in the power of the envelope (Fig. 1C, Urban Sounds). To compare the rhythmic structure of spoken language with that of urban sounds, we first compute the amplitude envelope for each sample in the training corpus of each dataset (English, Arabic, Bengali, Urban Sounds). We then calculate the power spectrum of each envelope and average them to obtain a representative modulation spectrum. The dataset for each language consists of spoken words from different speakers, while the dataset for urban sounds consists of common daily sounds, such as a jackhammer and car horn (Methods).

Rhythms in Recurrent Neural Networks

The regular temporal structure of speech raises the question of whether the auditory system encodes innate structural knowledge of the signal prior to language acquisition. What computational advantages could this provide? To investigate, we construct a network of coupled oscillators (Fig. 2) to model population-level activity in the auditory cortex [4, 11–14]. Each oscillator is defined by a characteristic frequency (rhythm) and

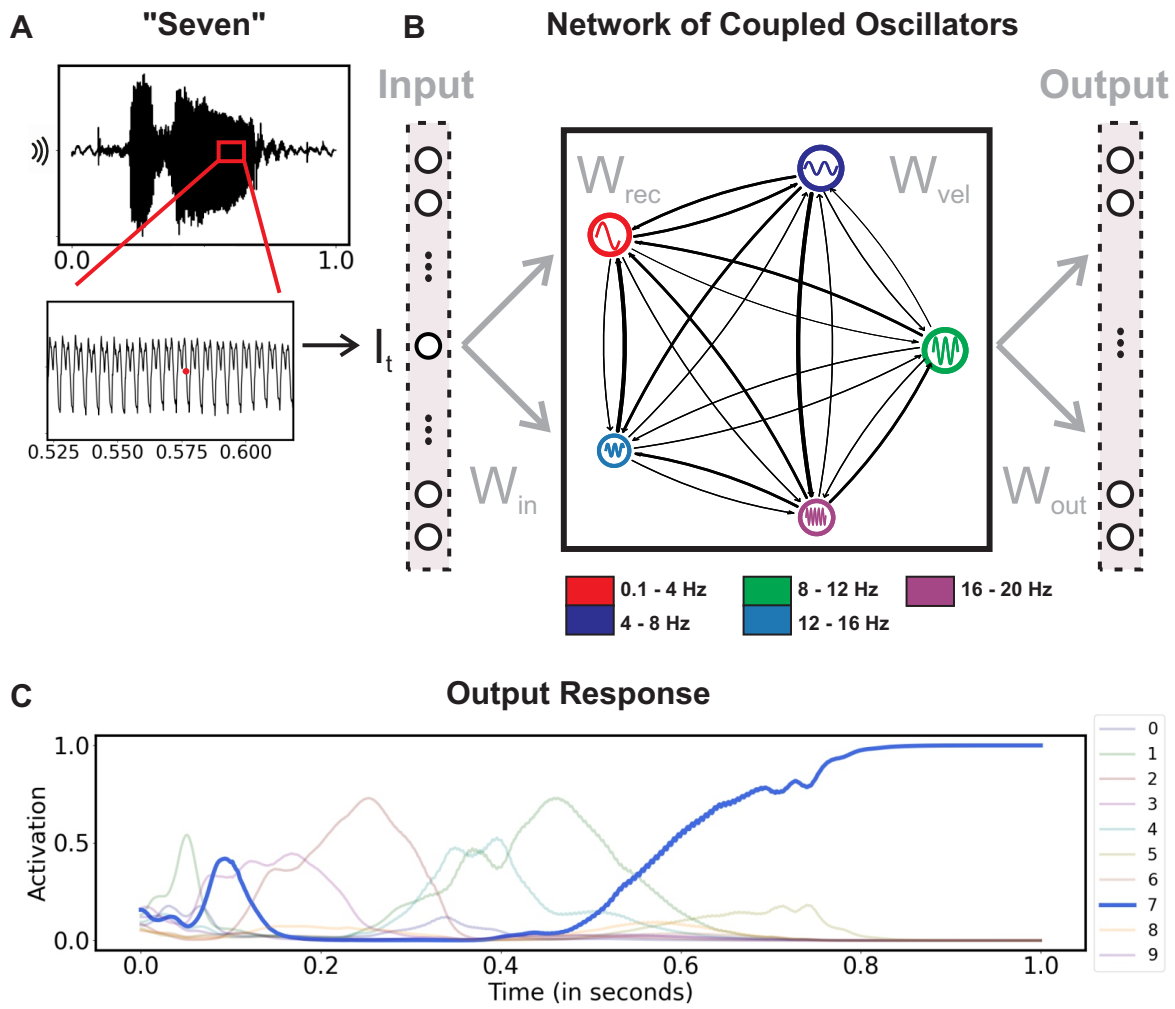

Figure 2: Network architecture. **(A)** Time-varying speech waveform of the spoken word "seven," sampled at each time point I_t. **(B)** Illustration of a network of coupled oscillators receiving samples of the speech waveform, with plastic input (W_{in}), recurrent (W_{rec}, W_{vel}), and output (W_{out}) weights. **(C)** Activity in the output layer in response to the spoken word.

a damping coefficient, which together determine its intrinsic response (Methods). The oscillators are simultaneously driven by this intrinsic dynamic and by raw speech input that modulate network activity. The network adopts an all-to-all coupling architecture, in which the activity of each oscillator influences every other unit, including itself. Internal couplings – determining how oscillators influence each other's activity – are initialized randomly, but adapt through training to encode stimulus structure. As a result, the network learns a rich and diverse set of trajectories in its state space, shaped by its intrinsic rhythms and the temporal regularities of speech.

We configure the oscillators with heterogeneous frequencies drawn from a fixed range. We evaluate several frequency ranges: some corresponding to brain oscillations (e.g., delta, theta, and gamma bands), while others having little correspondence to brain activity. During learning, we present raw auditory waveforms to the network, and the network

adapts its coupling parameters such that the waveforms – either speech or non-speech – are classified into appropriate categories (Methods). For instance, repeated presentations of the spoken word 'seven' guide the network toward a dynamical trajectory from which a linear readout layer determines the stimulus class. Once trained, we present the network with unseen samples of the learned categories, and record its recognition performance.

Performance Across Languages

To identify the effects of endogenous network oscillations on speech processing, we assess the performance of our model network in a series of tasks, including spoken word recognition in three different languages, and recognition of altered speech and non-speech sounds.

First, we initialize a network of 256 oscillators, with characteristic frequencies drawn uniformly at random from the range 0.1 Hz to 20 Hz. We train this network on a dataset of English spoken words, consisting of utterances of the words "zero" through "nine" by different speakers. Each spoken word in the training set was uttered multiple times. During training, we feed the network the raw speech waveform (as recorded from the speaker), and adjust the coupling strengths via the back-propagation through time (BPTT) learning algorithm (Methods) [15, 16]. After training the network for several epochs, we evaluate its word recognition performance on a held-out test set. As shown in Fig. 3A, the network achieves robust accuracy within the first few epochs, reaching a final performance of 89.20%($\pm$0.88%) after 50 epochs. Accuracy values are averaged over 10 random initializations of network parameters. Notably, this high test performance is achieved using only the raw speech waveform as input — without any pre-processing or feature extraction — during both training and testing.

For comparison, we train a recurrent neural network with `tanh` activation units – i.e. a network without endogenous oscillations – on the same dataset of spoken English digits, using the same learning algorithm and raw speech input. Unlike the oscillatory network, the `tanh`-based RNN was unable to learn the task (Fig. 3A). Test accuracy remained below chance level (4.90% $\pm$ 1.55%) throughout training.

Next, we assess which oscillator frequencies are most effective in recognizing speech. We initialize different networks, each configured with frequencies drawn uniformly at random from a specific range. We find that networks with oscillator frequencies in the range 1-5 Hz achieve the highest accuracy and training efficiency (Fig. 4A). Interestingly, this corresponds to the peak envelope power of the speech signal (Fig. 1C), indicating that a *match* between signal and network oscillations leads to the best accuracy. Accuracy dropped steeply for higher frequency network oscillations, falling to chance levels for frequencies above 40 Hz.

These results indicate that the coupled oscillatory network achieves high performance when it encodes knowledge of the temporal structure of speech – specifically, its amplitude modulation frequencies – prior to training. To assess whether this effect generalizes beyond English, we evaluate the network on two additional datasets: Arabic and Bengali spoken digits.

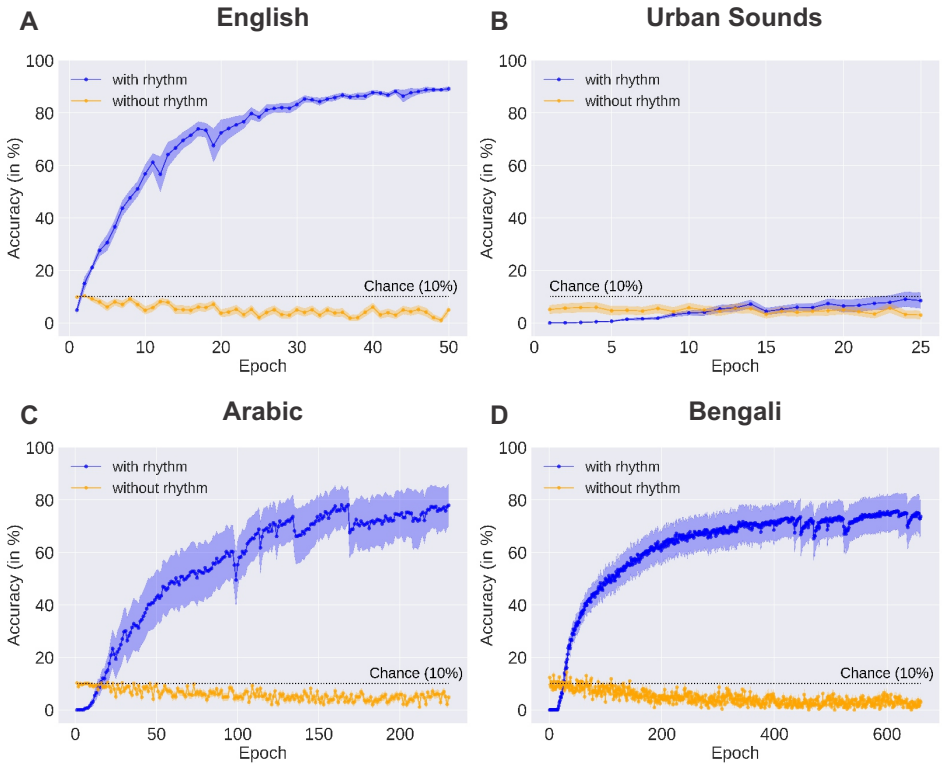

Figure 3: Performance with and without network rhythms. Average (test) accuracy per epoch of networks configured with oscillation frequencies chosen uniformly at random from the range [0.1 Hz, 20 Hz] (blue), compared to the same networks without oscillations (yellow). Shaded region depicts standard error across ten random seeds. Datasets were **(A)** AudioMNIST (English), **(B)** UrbanSound8k (Urban Sounds), **(C)** Arabic Spoken Digits, **(D)** Bengali Spoken Digits.

The Arabic spoken digits dataset contains utterances of the digits "zero" through "nine" in Arabic, each spoken multiple times by different speakers. As before, the network is trained using the raw speech waveform without any pre-processing. The results for Arabic closely mirror those observed for English, reinforcing the effectiveness of pre-configured low-frequency rhythms in speech processing. Networks configured with these low-frequency priors achieve substantially higher performance ($77.95\% \pm 8.05\%$) compared to non-rhythmic networks ($4.80\% \pm 1.52\%$, below chance; Fig. 3C). In contrast, networks configured with higher-frequency rhythms (> 20 Hz) are unable to learn, with performance dropping to chance levels (Fig. 4B), again showing that a match between signal and network frequencies is necessary for high accuracy.

The results for Bengali spoken digits confirm the same trend observed in English and Arabic, demonstrating the generality of this phenomenon. Networks configured with low-frequency rhythms achieved substantially higher performance ($74.20\% \pm 6.89\%$) than non-rhythmic networks ($3.01\% \pm 1.45\%$, below chance; Fig. 3D), with higher-frequency configurations again yielding chance-level accuracy for frequencies above 20 Hz (Fig. 4C).

Across all three datasets – English, Arabic, and Bengali – the results consistently demonstrate the effectiveness of pre-configured network rhythms in broadband speech

processing. Despite differences in phonetics, prosody, and speaker characteristics, low-frequency rhythms effectively capture the hierarchical temporal structure of the signals across these languages. This consistency suggests a general utility of these rhythms for processing natural speech in diverse linguistic contexts.

Performance for Non-Speech Sounds

Next, we assess whether the results for speech generalize to non-speech stimuli using a dataset of urban sounds [17]. Urban sounds, such as traffic or construction noise, lack the distinct temporal modulations characteristic of natural speech. We examine whether networks configured with low-frequency rhythms still provide a performance advantage. We find that those networks were no more effective than non-rhythmic networks in this task ($\sim$ 8.48% $\pm$ 2.99% and $\sim$ 2.11% $\pm$ 1.89%, respectively, both below chance level; Fig. 3B). This again suggests that endogenous network oscillations serve as prior knowledge of signal structure – the priors that are effective at speech recognition are entirely ineffective in recognizing signals that lack speech-like structure.

Speech rhythms can vary naturally across speakers and contexts, prompting the question of how such variability affects models with fixed temporal assumptions. In particular, we asked whether artificially altering the speech rate — by slowing it down (0.5$\times$) or speeding it up (2$\times$) — impacts network performance when using [1 Hz, 5 Hz] priors. We observe a significant drop in performance ($\sim$ 20%) when we either speed up or slow down the speech rate, compared to the natural speech rate (Supplementary Fig. S1). The observed performance drop reflects a misalignment between the network's intrinsic priors and the altered speech rates.

State-Space Dynamics

To investigate how dynamical priors shape the representations in the network, we plot the first three principal components of network activity in response to the spoken words. We find that before training, network trajectories across words and languages are almost identical, with only minor variations reflecting subtleties in intonation and stress patterns (Fig. 5A-C: Untrained). During training, the trajectories become increasingly distinct for each spoken word, and once training is complete, the trajectories are highly decorrelated, encoding the specific patterns of each word across the three languages (Fig. 5A-C: Trained). In contrast, networks that lack intrinsic rhythms do not learn decorrelated representations across words and languages (Fig. 6). The scaffold provided by the low-frequency rhythms, therefore, enable class-wise trajectory disentanglement via training, which is the reason underlying high classification accuracy (Fig. 3).

We then assess which oscillators predominantly shape activity when a network is initialized with frequencies drawn uniformly at random from the range [0.1 Hz, 20 Hz]. For this, we calculate the component loadings of the first principal component of the network's activity after training. The first principle component captures $\sim$ 70% of the total variance. Across the three languages, we observe that oscillators within 0.1 Hz and 2.5 Hz dominate the network's activity (Fig. 7), with higher-frequency oscillators ($>$ 5 Hz) exerting only a weak influence. Thus, consistent with our other results, low-frequency oscillations were the primary drivers of speech computations.

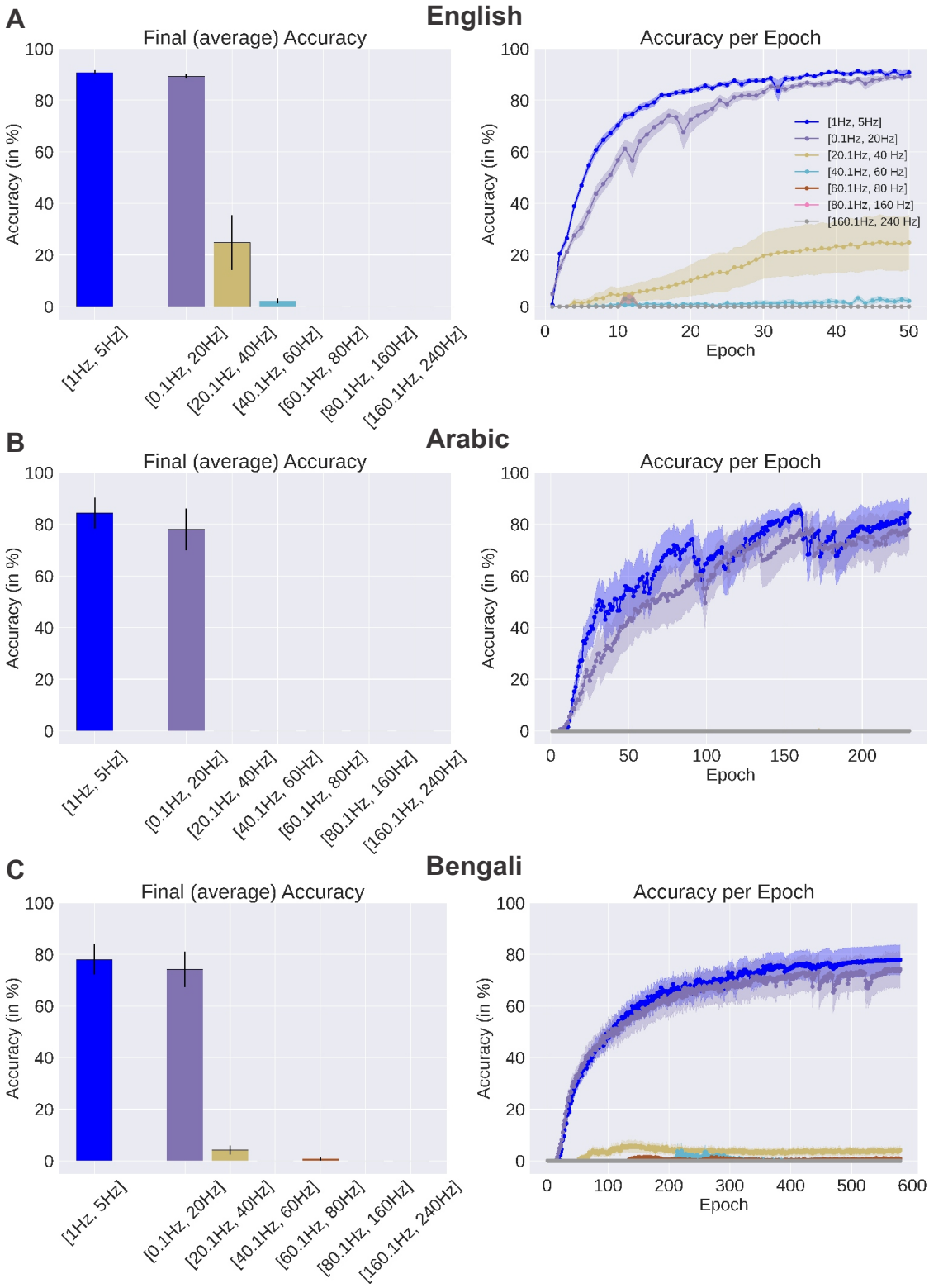

Figure 4: Performance across rhythms. **(A)** Accuracy in detecting English spoken words. Left: Average test accuracy of networks configured with oscillator frequencies drawn uniformly at random from the specified ranges. Values are averaged across ten random seeds. Error bars depict standard error. Performance is highest for low-frequency oscillations that match the temporal structure of the signal. Right: Average (test) accuracy per epoch of the same networks. **(B, C)** Arabic, Bengali

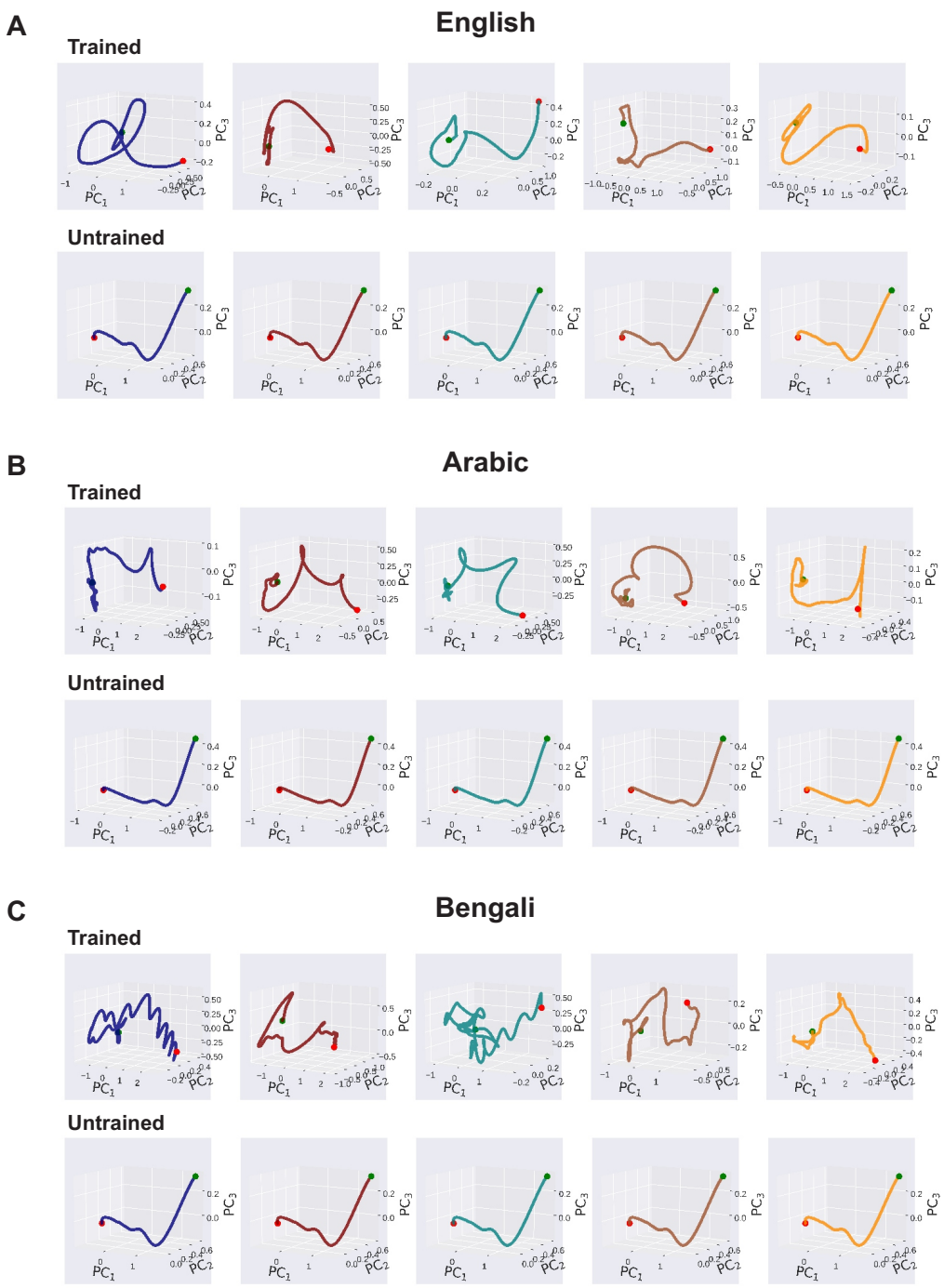

Figure 5: Neural trajectories with endogenous oscillations. **(A)** Trajectory of network activity in PCA space in response to five spoken digits in English ("zero", "two", "four", "six", "eight"). Trajectories are averaged across all test samples in the dataset. The network is initialized with frequencies drawn uniformly at random from the range [0.1 Hz, 20 Hz]. Green and red dots depict start and end of the trajectories respectively. Top: before training. Bottom: after training. **(B, C)** Arabic, Bengali.

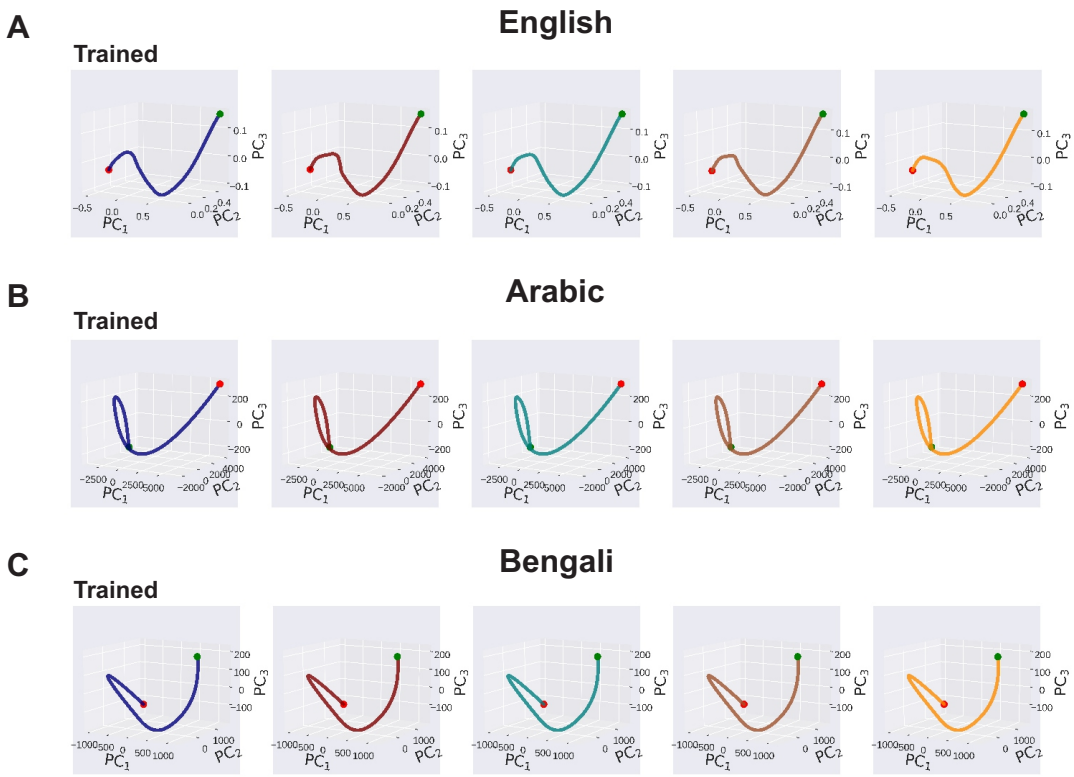

Figure 6: Neural trajectories without endogenous oscillations. **(A)** Trajectory of neural activity in PCA space in response to five spoken digits in English ("zero", "two", "four", "six", "eight") for a network with no endogenous oscillations and randomly initialized weights. Different digits evoke the same trajectory. Trajectories are averaged across all test samples in the dataset. Green and red dots depict start and end of the trajectories respectively. **(B, C)** Arabic, Bengali.

Discussion

We show that networks of coupled oscillators with endogenous low-frequency oscillations accurately learn to recognize spoken words in three distinct languages from raw audio input. Our results demonstrate that oscillations serve as prior knowledge of speech structure, providing a temporal scaffold on which statistical learning can take place. Unit oscillations drawn from the 1-5 Hz frequency range results in the best task performance in all three languages. This suggests that despite differences in phonetic and prosodic features of English, Arabic, and Bengali, their raw acoustic signals share common, perhaps universal, temporal structure. It has been proposed that speech production involves quasi-rhythmic motor patterns dictated by skeletal, muscular, and neural circuit-level constraints of the human speech-motor system [4, 18, 19]. The perceptual system, in turn, imposes rhythms that are tuned to detect and interpret speech signals at the appropriate temporal granularity, akin to active sensing in other modalities, but without explicit movement-driven sampling. Our results lend computational support to this view.

Imposing oscillatory dynamics in itself is insufficient for extracting task relevant structure. Rather, a match between signal and network oscillations is essential. We find that networks with intrinsic oscillations mirroring those of the audio input reach high speech

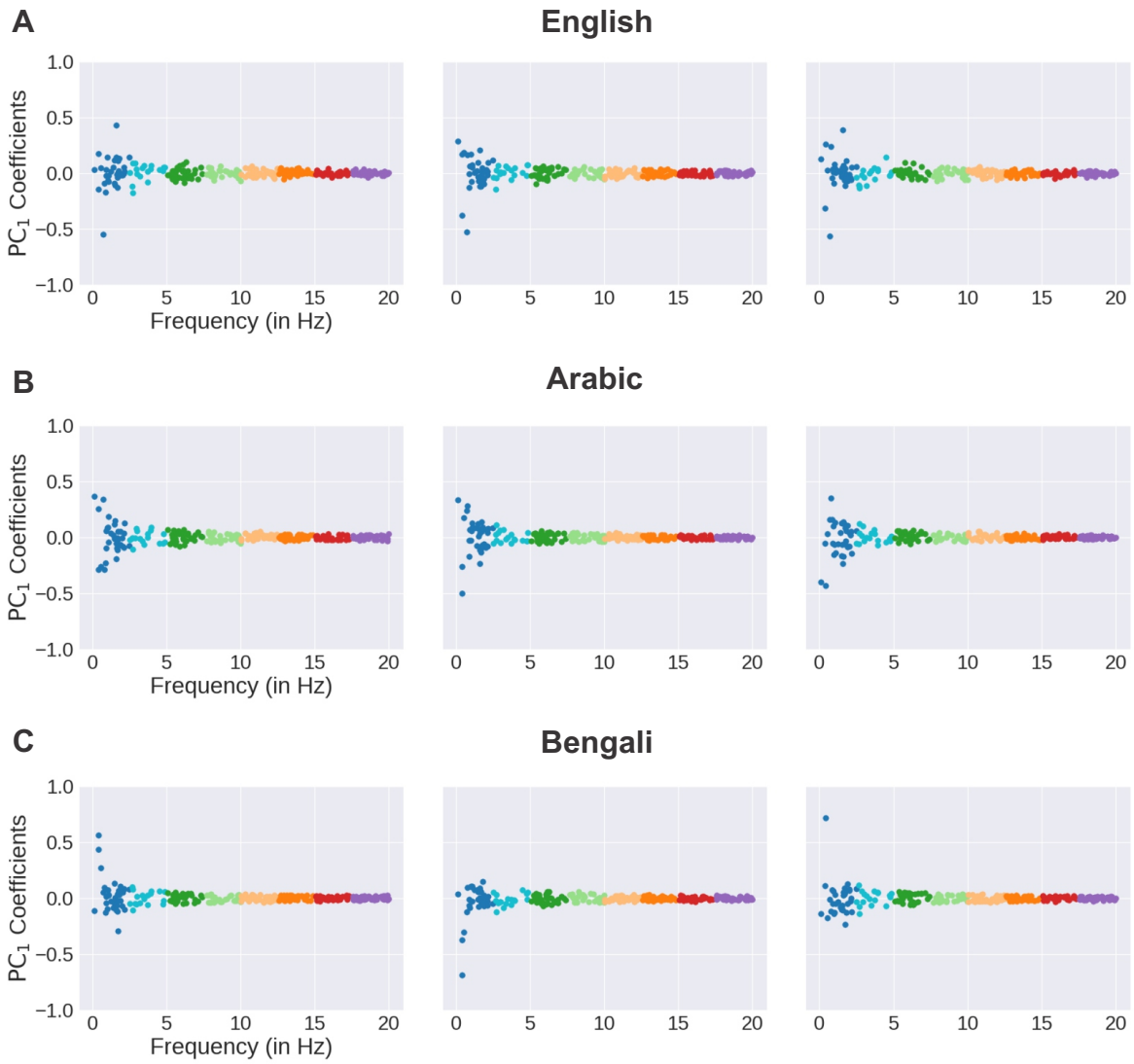

Figure 7: Dominant oscillation mode. The influence of each oscillator to the first principal component in response to **(A)** English, **(B)** Arabic, and **(C)** Bengali spoken words for a network trained with randomly initialized frequencies from the range [0.1 Hz, 20 Hz].

recognition accuracy, whereas networks with oscillations mismatched to the signal are unable to learn the task. In circuit theory, frequency matching between system and input, or *resonance*, imparts frequency selectivity, voltage magnification, and other signal transformations, widely used in analog communication systems. In analogy, circuit resonance in the brain has been reported in several studies, where signals arriving in phase with the natural oscillations of cortical circuits are selectively gated or amplified [20–22]. In our model, we use an explicit parameter γ to define the characteristic frequency of each node, and suggest it be interpreted as a *dynamic prior* that enables effective categorization of signals with matching temporal structure. We note that although the initial resonance properties of each unit is determined by γ, the learned frequency response is more complex due to network influences, as also reported in other studies [11, 12].

In machine learning, priors, or inductive biases, refer to pre-existing knowledge or initial assumptions given to a model as cues to guide the learning process. This enables faster and more efficient training. A prior may be encoded by imposing a particular structure on the form of a model or as a hint about the underlying distribution of the input (e.g., sparsity regularization in LASSO). Whether explicit or implicit, priors lend direction to the learning algorithm to find meaningful patterns in the data, narrowing the range of possible hypotheses about the class of functions being learned [23]. In a neural network, finding good minima in a high-dimensional, non-convex loss landscape can be difficult [24]. Priors help circumvent this difficulty by initializing the network to favorable configurations, enabling it to avoid adverse terrain of the loss landscape.

In the brain, priors come encoded at birth, though their precise forms are not well understood. They may take the form of stereotyped circuit architectures imposed by development [25–28], or as we propose here, particular types of dynamical response. Learning involves adapting these initial expectations to the patterns of stimuli encountered in the world. For example, in language acquisition, priors may encode general syntactic constraints, with exposure to a particular language resulting in language-specific expectations and associations with sound. The learning rules in the brain likely differ substantially from the BPTT algorithm used here [29–32], although alternative forms of credit assignment may be partly at play [33–35].

Network structure in the brain has inspired inductive biases in artificial neural networks [36–38]. The topographic organization of the primary visual cortex, for example, where neurons respond selectively to localized regions of the visual field, has motivated the use of local filters in convolutional neural networks. By enforcing spatial locality and translational invariance, these architectural choices encode knowledge essential for efficient processing of visual information [39, 40]. Beyond this spatial organization, neural systems exhibit rich temporal dynamics, often exhibiting traveling waves of activity patterns that unfold across the cortical surface. When embedded within recurrent neural networks (RNNs), traveling waves generate dynamic representations that can be exploited for efficient sequence learning, continual learning, and long-term memory consolidation [41–46]. Other phenomena characteristic of coupled neural systems, including transient synchronization, phase locking, and the hierarchical nesting of oscillations, may also be incorporated into similar wave-based models. These wave-RNNs offer new tools to artificial intelligence and serve as a platform for testing computational hypotheses about brain function.

Acknowledgments

This work was supported by NSF grants 2319060 and 2223811 to NI.

References

[1] Anne-Lise Giraud et al. "Endogenous cortical rhythms determine cerebral specialization for speech perception and production". *Neuron* 56.6 (2007), pp. 1127–1134.

[2] Huan Luo, Zuxiang Liu, and David Poeppel. "Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation". *PLoS Biology* 8.8 (2010), e1000445.

[3] Maria Clemencia Ortiz-Barajas, Ramón Guevara, and Judit Gervain. "Neural oscillations and speech processing at birth". *Iscience* 26.11 (2023).

[4] David Poeppel and M Florencia Assaneo. "Speech rhythms and their neural foundations". *Nature Reviews Neuroscience* 21.6 (2020), pp. 322–334.

[5] Stuart Rosen. "Temporal information in speech: acoustic, auditory and linguistic aspects". *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 336.1278 (1992), pp. 367–373.

[6] Nai Ding et al. "Temporal modulations in speech and music". *Neuroscience & Biobehavioral Reviews* 81 (2017), pp. 181–187.

[7] Oded Ghitza and Steven Greenberg. "On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence". *Phonetica* 66.1-2 (2009), pp. 113–126.

[8] György Buzsáki and Mihály Vöröslakos. "Brain rhythms have come of age". *Neuron* 111.7 (2023), pp. 922–926.

[9] Frank C Hoppensteadt and Eugene M Izhikevich. *Weakly connected neural networks*. Vol. 126. Springer Science & Business Media, 2012.

[10] Hugh R Wilson and Jack D Cowan. "Excitatory and inhibitory interactions in localized populations of model neurons". *Biophysical Journal* 12.1 (1972), pp. 1–24.

[11] Felix Effenberger et al. "The functional role of oscillatory dynamics in neocortical circuits: a computational perspective". *Proceedings of the National Academy of Sciences* 122.4 (2025), e2412830122.

[12] T Konstantin Rusch and Siddhartha Mishra. "Coupled oscillatory recurrent neural network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies". *International Conference on Learning Representations* (2021).

[13] Anne-Lise Giraud and David Poeppel. "Cortical oscillations and speech processing: emerging computational principles and operations". *Nature Neuroscience* 15.4 (2012), pp. 511–517.

[14] Oded Ghitza. "Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm". *Frontiers in Psychology* 2 (2011), p. 130.

[15] Paul J Werbos. “Backpropagation through time: what it does and how to do it”. *Proceedings of the IEEE* 78.10 (1990), pp. 1550–1560.

[16] Ian Goodfellow et al. *Deep learning*. Vol. 1. 2. MIT Press, 2016.

[17] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. “A dataset and taxonomy for urban sound research”. *ACM Multimedia* (2014).

[18] Frank H Guenther. *Neural control of speech*. MIT Press, 2016.

[19] Peter F MacNeilage. “The frame/content theory of evolution of speech production”. *Behavioral and brain sciences* 21.4 (1998), pp. 499–511.

[20] Bruce W Knight. “The relationship between the firing rate of a single neuron and the level of activity in a population of neurons: Experimental evidence for resonant enhancement in the population response”. *The Journal of General Physiology* 59.6 (1972), pp. 767–778.

[21] Eugene M Izhikevich. “Resonate-and-fire neurons”. *Neural Networks* 14.6-7 (2001), pp. 883–894.

[22] Bruce Hutcheon and Yosef Yarom. “Resonance, oscillation and the intrinsic frequency preferences of neurons”. *Trends in Neurosciences* 23.5 (2000), pp. 216–222.

[23] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[24] Hao Li et al. “Visualizing the loss landscape of neural nets”. *Advances in Neural Information Processing Systems* 31 (2018).

[25] Nabil Imam et al. “Dual Computational Systems in the Development and Evolution of Mammalian Brains”. *bioRxiv* (2024), pp. 2024–11.

[26] Nabil Imam and Barbara L. Finlay. “Self-organization of cortical areas in the development and evolution of neocortex”. *Proceedings of the National Academy of Sciences* 117.46 (2020), pp. 29212–29220.

[27] Anthony M Zador. “A critique of pure learning and what artificial neural networks can learn from animal brains”. *Nature Communications* 10.1 (2019), p. 3770.

[28] Sergey Shuvaev et al. “Encoding innate ability through a genomic bottleneck”. *Proceedings of the National Academy of Sciences* 121.38 (2024), e2409160121.

[29] Daniel E Feldman. “The spike-timing dependence of plasticity”. *Neuron* 75.4 (2012), pp. 556–571.

[30] Cengiz Pehlevan, Anirvan M Sengupta, and Dmitri B Chklovskii. “Why do similarity matching objectives lead to Hebbian/anti-Hebbian networks?” *Neural Computation* 30.1 (2017), pp. 84–124.

[31] Christiane Linster and Thomas A Cleland. “Decorrelation of odor representations via spike timing-dependent plasticity”. *Frontiers in Computational Neuroscience* 4 (2010), p. 157.

[32] Jason J Moore et al. “The neuron as a direct data-driven controller”. *Proceedings of the National Academy of Sciences* 121.27 (2024), e2311893121.

[33] Robert Gütig. “Spiking neurons can discover predictive features by aggregate-label learning”. *Science* 351.6277 (2016), aab4113.

[34] Guillaume Bellec et al. “A solution to the learning dilemma for recurrent networks of spiking neurons”. *Nature Communications* 11.1 (2020), p. 3625.

[35] Yuhan Helena Liu et al. “Cell-type–specific neuromodulation guides synaptic credit assignment in a spiking neural network”. *Proceedings of the National Academy of Sciences* 118.51 (2021), e2111821118.

[36] Yann LeCun et al. “Backpropagation applied to handwritten zip code recognition”. *Neural Computation* 1.4 (1989), pp. 541–551.

[37] Jeffrey L Elman. “Finding structure in time”. *Cognitive Science* 14.2 (1990), pp. 179–211.

[38] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. *Nature* 521.7553 (2015), pp. 436–444.

[39] David J Field. “What is the goal of sensory coding?” *Neural Computation* 6.4 (1994), pp. 559–601.

[40] Mayukh Deb, Mainak Deb, and N Murty. “TopoNets: High performing vision and language models with brain-like topography”. *arXiv preprint arXiv:2501.16396* (2025).

[41] T Anderson Keller and Max Welling. “Neural wave machines: learning spatiotemporally structured representations with locally coupled oscillatory recurrent neural networks”. *International Conference on Machine Learning* (2023).

[42] Samuel Lanthaler, T Konstantin Rusch, and Siddhartha Mishra. “Neural oscillators are universal”. *Advances in Neural Information Processing Systems* 36 (2023), pp. 46786–46806.

[43] Lyle Muller, Patricia S Churchland, and Terrence J Sejnowski. “Transformers and cortical waves: encoders for pulling in context across time”. *Trends in Neurosciences* (2024).

[44] T. Anderson Keller et al. “Traveling Waves Encode the Recent Past and Enhance Sequence Learning”. *International Conference on Learning Representations (ICLR)* (2024).

[45] Timothy Tadros et al. “Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks”. *Nature Communications* 13.1 (2022), p. 7742.

[46] Lyle Muller et al. “Cortical travelling waves: mechanisms and computational principles”. *Nature Reviews Neuroscience* 19.5 (2018), pp. 255–268.

[47] Sören Becker et al. “AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark”. *Journal of the Franklin Institute* 361.1 (2024), pp. 418–428.

[48] Abdulrahman Alalshekmubarak and Leslie S. Smith. “On improving the classification capability of reservoir computing for Arabic speech recognition”. *International Conference on Artificial Neural Networks* (2014).

[49] Mir Sayeed Mohammad, Azizul Zahid, and Md Asif Iqbal. “BanglaNum–A Public Dataset for Bengali Digit Recognition from Speech”. *arXiv preprint arXiv:2403.13465* (2024).

Methods

Datasets. We used the AudioMNIST (English), Arabic Spoken Digits, BanglaNum (Bengali), and UrbanSound8K datasets in our tasks. AudioMNIST consists of 30,000 audio recordings ($\sim$ 8.3 hours) of spoken digits (0 through 9) across 60 different English speakers [47]. Each recording is $\sim$ 1 second, originally sampled at 48 kHz on a single-channel. Each speaker recited each digit 50 times, resulting in 3,000 recordings across each class.

The Arabic Spoken Digits dataset is a subset of the Arabic Speech Corpus for Isolated Words [48]. The dataset consists of approximately $5,000$ audio recordings of spoken digits (0 through 9) across 20 different Arabic speakers. Each recording is $\sim$ 1 second, originally sampled at $44,100$ Hz on two channels (i.e., stereo). The Bengali Spoken Digits (BanglaNum) dataset consists of approximately $2,250$ audio recordings of spoken digits (0 through 9) across 40 different Bengali speakers [49]. Each recording is $\sim$ 0.5 seconds, originally sampled at 16 kHz on a standard single-channel microphone.

UrbanSound8K consists of field recordings of various sounds found in an urban environment, spanning 10 low-level classes (i.e., air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music) [17]. In total, there are 8,732 audio recordings ($\sim$ 7.06 hours), each with a variable time length ($\leq$ 4 seconds) with sampling rates varying from 8 kHz to 96 kHz.

For the language datasets, we randomly select 90% of the samples to be used as the training set, and the rest as the test set. For Urban Sounds, we randomly select 80% as the training set, and the rest as the test set. For all datasets, we first down sample each recording to 8 kHz, and then normalize the amplitude of the signal to the interval $[-1, 1]$. The resulting time-series, $I(t)$, is used as the input signal to the network.

Model Description. We use the coRNN (Coupled Oscillatory Recurrent Neural Network) setup, which is an RNN architecture of time-discretized coupled oscillators [12]. In the coRNN, the coupled oscillators take the form of the following second-order, m-dimensional system of ODEs:

$$\mathbf{y}''(t) = \sigma(W^{Rec}\mathbf{y}(t) + W^{Vel}\mathbf{y}'(t) + W^{in}\mathbf{I}(t) + \mathbf{b}) - \gamma\mathbf{y}(t) - \epsilon\mathbf{y}'(t) \tag{1}$$

where,

$\mathbf{I}(t) \in \mathbb{R}^d$	time-varying input signal
$W^{\text{in}} \in \mathbb{R}^{m\times d}$	input weight matrix
$\mathbf{y}(t) \in \mathbb{R}^m$	hidden states of the RNN
$\mathbf{y}'(t) \in \mathbb{R}^m$	velocity of hidden states of the RNN
$W^{\text{Rec}}, W^{\text{Vel}} \in \mathbb{R}^{m\times m}$	recurrent weight matrices
$\mathbf{b} \in \mathbb{R}^m$	bias term
$\sigma(\cdot) \in \mathbb{R}$	activation function (tanh)
$\boldsymbol{\gamma} \in \mathbb{R}^m$	characteristic frequency (Hz) of each node (≥ 0)
$\boldsymbol{\epsilon} \in \mathbb{R}^m$	damping of each node (≥ 0)

Each node in the network is parameterized by a fixed characteristic frequency (γ) and damping (ϵ) coefficient. The weight matrices, W^{Rec}, W^{Vel}, & W^{in}, represent the *coupling*

& *damping* between oscillators, and the stimulus-driven term, respectively. In Eq. 1, time, $t \in [0, 1]$, represents a continuous variable. In practice, to simulate the equation, we let the velocity term, $\mathbf{y}'(t) = \mathbf{z}(t)$. We can then rewrite the second-order ODE (Eq. 1) as a first-order ODE (Eq. 2), where $\mathbf{z}'(t) = \mathbf{y}''(t)$, resulting in:

$$\mathbf{z}'(t) = \sigma(W^{Rec}\mathbf{y}(t) + W^{Vel}\mathbf{z}(t) + W^{in}\mathbf{I}(t) + \mathbf{b}) - \gamma\mathbf{y}(t) - \epsilon\mathbf{z}(t) \tag{2}$$

For a fixed timestep, $0 < \Delta t < 1$, the IMEX (implicit-explicit) discretization of this equation results in:

$$\mathbf{y}_t = \mathbf{y}_{t-1} + \mathbf{z}_t \Delta t \tag{3}$$

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \Delta t\sigma(W^{Rec}\mathbf{y}_{t-1} + W^{Vel}\mathbf{z}_{t-1} + W^{in}I_t + \mathbf{b}) - \Delta t\gamma\mathbf{y}_{t-1} - \Delta t\epsilon\mathbf{z}_{t-1} \tag{4}$$

We consider only the explicit discretization of equations 3 and 4 in the network simulations.

Simulations. We simulate the coRNN both with and without endogenous frequencies, using a network of 256 units in each case. In the endogenous case, we configure the network with various combinations of oscillations and damping (Table S1 in Supplementary Information). In the case without endogenous frequencies, we set γ and ϵ to 0 in Eq. 1, effectively resulting in an Elman (or standard) RNN [37].

All simulated networks consist of a trainable input, recurrent, and readout (SoftMax) layer, where the initial weights are sampled from a uniform distribution, $W \sim \mathcal{U}(-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}})$ (n is the input dimension to each layer). The input to each network, $I(t)$, is a time-varying signal. All parameters were subsequently trained via BPTT (Backpropagation Through Time). We used the cross-entropy loss function with the AdamW optimizer, a learning rate of $4.2 \cdot 10^{-3}$, and a time step of $5 \cdot 10^{-3}$.

For the AudioMNIST, Arabic Spoken Digits, Bengali Spoken Digits dataset, we used 50 training epochs with a batch size of 64, which took a total of $\sim$ 50 hours of training time per network on a NVIDIA V100 GPU. For the UrbanSounds8k dataset, we used 25 epochs with a batch size of 32, which took $\sim$ 25 hours to train. We carried out a systematic scan of the frequency and damping ranges (Table S1 in Supplementary Information), but did not perform an exhaustive grid-search due to the long training times.

To assess network performance, we used the following classification scheme: for the last $\sim$ 62.5 ms of the input signal, we take the node in the SoftMax layer with the highest activation. Specifically, if the maximum SoftMax activation at time t exceed a threshold of 0.60, the corresponding class i is assigned as the prediction at that time point; otherwise, the output is labeled as Null. After collecting predictions across the classification window, we determined the final class label as follows: if class i is the most frequently predicted label in this window and appears in at least 80% of the non-null predictions, then i is selected as the network's final decision. Otherwise, the classification is marked as uncertain. This scheme ensures a robust classification at readout, despite the oscillatory nature of the network. See Fig. 2C for an illustration of a network response.

Inductive Biases. Here we discuss the interaction between the frequency (γ) and damping (ϵ) terms in Eq. 1. Although these hyperparameters remain fixed, the coupling of the frequency (W^{rec}) and damping (W^{vel}) between nodes is learned. The choice of the frequency and damping of each node biases the learning process toward certain oscillations. The learned coupling allows sufficient flexibility, allowing the network to express solutions that emerge from network interactions.

Because of global coupling in the network, the intrinsic frequency and damping of each node influences the oscillation that emerges in other nodes. An oscillator may be pulled toward a certain oscillatory state, given the influence of the frequency and damping of other oscillators. Strong priors help circumvent the deviation of an oscillator from its natural frequency.

Consider, a network of coupled oscillators without intrinsic frequencies, but with fixed damping; that is, we set $\boldsymbol{\gamma} = \mathbf{0}$ in Eq. 1, while setting $\boldsymbol{\epsilon}$ to a fixed positive value. In this setup, the frequency parameter has no influence on the oscillation of a node – the node will oscillate based on the learned (input and recurrent) coupling and the damping parameters. Although the span of oscillations the node can assume is much larger (compared to a network with non-zero $\boldsymbol{\gamma}$), it may take much longer to converge on an effective oscillatory solution. The benefits of the co-RNN come from imposing the right frequency priors before learning.

In our network simulations, we explore the consequence of no frequency prior, but fixed damping range, [0.1, 80]. Initially, we observe that these networks struggle to learn the task, but with enough training can eventually learn to oscillate at ~ 1 Hz. In contrast, networks with ~ 1 Hz priors learn the task rapidly. An appropriate scaffold is thus essential for the learning process.

Supplementary Information

Frequency Range	Damping Range	English	Arabic	Bengali
[1, 5] Hz	[0.1, 80]	**90.79%** ± **0.84%**	**84.28%** ± **6.08%**	**78.01%** ± **5.82%**
[0.1, 20] Hz	[0.1, 80]	89.20% ± 0.88%	77.95% ± 8.05%	74.20% ± 6.89%
[20.1, 40] Hz	[20.1, 40]	24.81% ± 10.60%	0.00% ± 0.00%	4.15% ± 1.77%
[40.1, 60] Hz	[40.1, 60]	2.17% ± 0.90%	0.00% ± 0.00%	0.00% ± 0.00%
[60.1, 80] Hz	[60.1, 80]	0.00% ± 0.00%	0.00% ± 0.00%	0.73% ± 0.60%
[80.1, 160] Hz	[80.1, 160]	0.00% ± 0.00%	0.00% ± 0.00%	0.00% ± 0.00%
[160.1, 240] Hz	[160.1, 240]	0.00% ± 0.00%	0.00% ± 0.00%	0.00% ± 0.00%
[250.1, 330] Hz	[250.1, 330]	0.00% ± 0.00%	0.00% ± 0.00%	0.00% ± 0.00%
[400.1, 480] Hz	[400.1, 480]	0.00% ± 0.00%	0.00% ± 0.00%	0.00% ± 0.00%
[1000.1, 1080] Hz	[1000.1, 1080]	0.00% ± 0.00%	0.00% ± 0.00%	0.00% ± 0.00%
No Rhythm	No Damping	4.90% ± 1.55%	4.80% ± 1.52%	3.01% ± 1.45%

Table S1: Final test accuracy of networks configured with frequency and damping selected from the specified ranges. The bold values indicate the highest accuracies. Values are averaged (± standard error) across 10 random seeds.

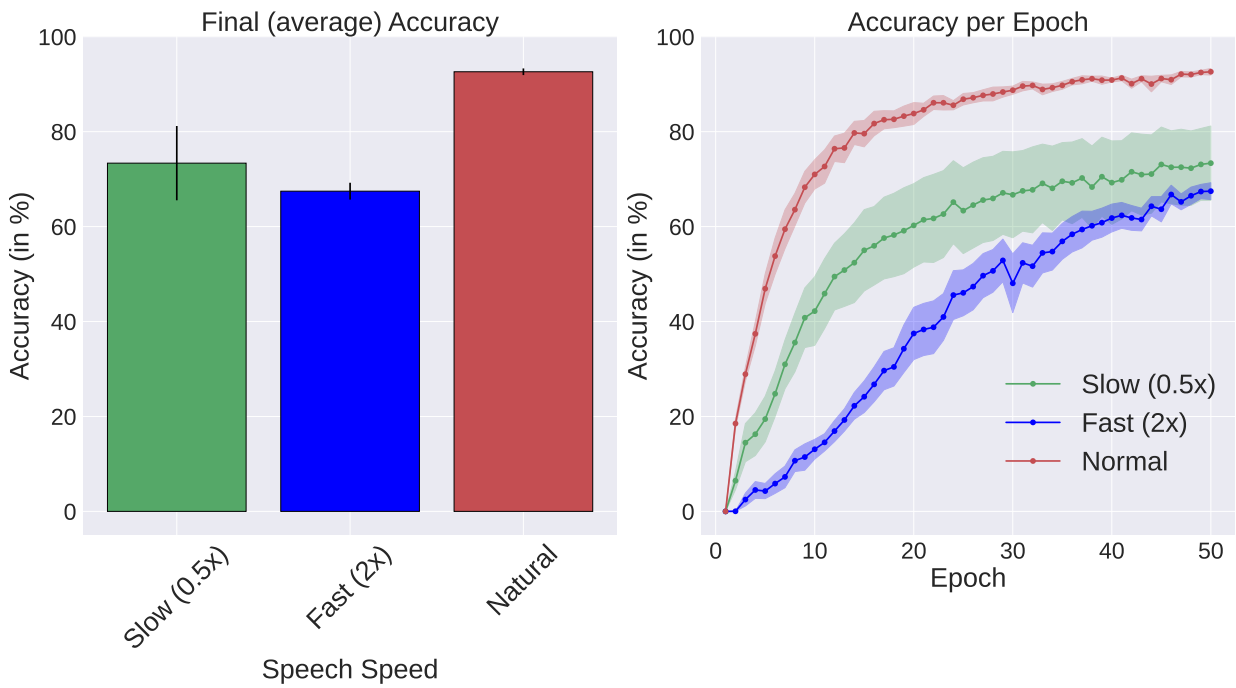

Figure S1: *Left:* Final (test) performance of networks configured with [1 Hz, 5 Hz] endogenous frequencies when the speech signal is slowed ($0.5x$), sped up ($2x$), and natural. *Right:* Average (test) accuracy per epoch for the same networks. The shaded region depicts standard error. A total of ten random seeds is used.