Predictive Dynamics Improve Noise Robustness in a Deep Network Model of the Human Auditory System

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sensory systems are robust to many types of corrupting noise. However, the neural mechanisms that drive robustness are unclear. Empirical evidence suggests that top-down predictions are important for processing noisy stimuli, and the substantial feedback connections in primate sensory cortices have been proposed to facilitate these predictions. Here, we implement predictive dynamics in a large scale model of the human auditory system. Specifically, we augment a feedforward deep neural network trained on noisy speech classification with a recently introduced predictive feedback scheme. We find that predictive dynamics improve speech identification across several types of corrupting noise. These performance gains were associated with denoising of network representations and alterations in layer dimensionality. Finally, we find that the model captures brain data outside of the speech domain. Overall, this work demonstrates that predictive dynamics are a candidate mechanism for human auditory robustness and provides a testbed for hypotheses regarding the dynamics of auditory representations. Additionally, we discuss the potential for this framework to provide insight into robustness mechanisms across sensory modalities.

1 Introduction

Our ability to process complex sensory information requires the isolation of stimuli of interest from background noise. For instance, we can pick out a friend's voice in a crowded coffee shop or spot a predator in a grassy field. To distinguish these stimuli of interest, your sensory system must somehow become robust to background noise. A long standing goal in the field of sensory neuroscience is to understand the neural mechanisms that enable human sensory systems to solve these problems [1–3].

A large body of experimental evidence suggests that top-down predictions play an important role in processing stimuli in noisy contexts [4–6]. Feedback connections are abundant in sensory cortices [7, 8], and have been hypothesized to carry predictive signals [9] but are often omitted in large scale models. In the visual neuroscience literature, imbuing computational models with biologically-inspired predictive dynamics has been found to increase performance in noisy object recognition tasks [10–12]. These insights offer a promising indication that predictive dynamics play a role in enabling noise robustness.

Here, we explore noise robustness in the context of auditory perception. We augmented a deep neural network of the human auditory system with predictive dynamics and explored how these dynamics affected network representations and performance. We find that predictive dynamics improve robustness across several types of real-world noise. We explore different mechanisms that support this robustness. Finally, we discuss how the intrinsically temporal nature of auditory

Submitted to 4th Workshop on Shared Visual Representations in Human and Machine Visual Intelligence (SVRHM) at NeurIPS 2022. Do not distribute.

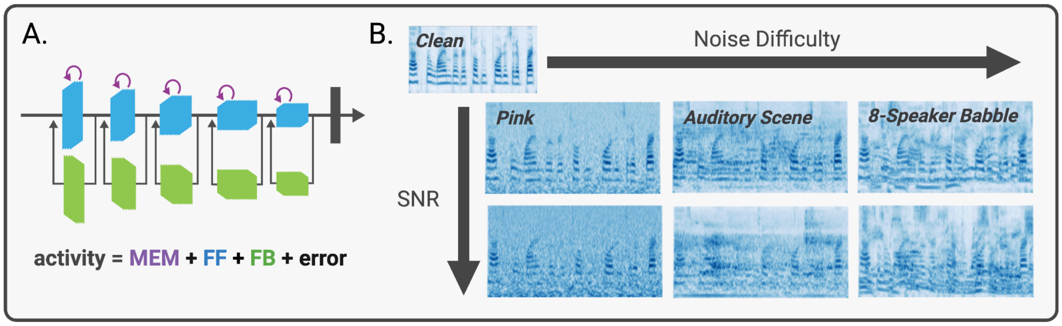

Figure 1: **A.** Schematic of network and predictive dynamics. FF: feedforward, FB: feedback, ER: error correction, MEM: memory. Each block indicates a convolutional layer. Error is not depicted in the diagram. **(B)** Clean sounds were embedded in three background noises of varying difficulty, with different grades of SNR.

perception lends itself to interesting sensory perception problems that are general across many modalities.

2 Model

2.1 Predictive dynamics in a deep neural network are introduced with the Predify framework.

We use the Predify framework [10] to incorporate predictive dynamics into a feedforward network. Briefly, given input into the network at time 0, the activity of each layer i at some time t is a linear combination of four terms (Fig 1A; see Appendix A.1): **(1)** Feedforward drive taking activity from layer $i - 1$ as input, **(2)** Feedback from layer $i + 1$ that predicts the activity at layer i, **(3)** Error correction that computes the gradient of the prediction error of layer i to correct the activity at layer i, **(4)** Memory that integrates over the activity from time $t - 1$.

Under this framework, higher-order regions predict and correct the activity of lower-order regions. Each layer has a set of hyperparameters that control the proportional strength of each of these terms. Note that hyperparameters for feedforward, feedback, and memory are constrained to sum to 1. Different settings of the hyperparameters can capture a variety of known dynamics, from purely feedforward to classical predictive coding [10].

2.2 Feedforward weights are learned in a supervised auditory classification task

We use the feedforward neural network introduced by [13] to define the feedforward weights of the predictive network. This network was trained to classify words given cochleagram inputs (time-frequency decomposition of sounds modeled after the human cochlea, Appendix A.2). Inputs were embedded in real-world background noise of varying signal-to-noise ratio (Fig 1B).

We evaluate the performance of the network on speech classification with speech embedded in pink noise, auditory scene, and 8-speaker babble (Fig 1B). These backgrounds were chosen because humans show variable speech recognition performance across these conditions ([13]).

This network was chosen because of its behavioral similarity to humans and ability to capture human auditory cortical responses more accurately than other models [13].

2.3 Feedback weights are learned in an unsupervised reconstruction task

Consistent with [10], we froze the feedforward weights and trained the feedback connections in an unsupervised fashion. Each feedback layer i was trained to reconstruct the activity at layer $i - 1$ given the activity at layer i. Importantly, the inputs used for this training step have no background noise.

Finally, the hyperparameters are optimized as in [12]. Optimization was performed for each combination of background noise and SNR to maximize classification performance on the task from [13].

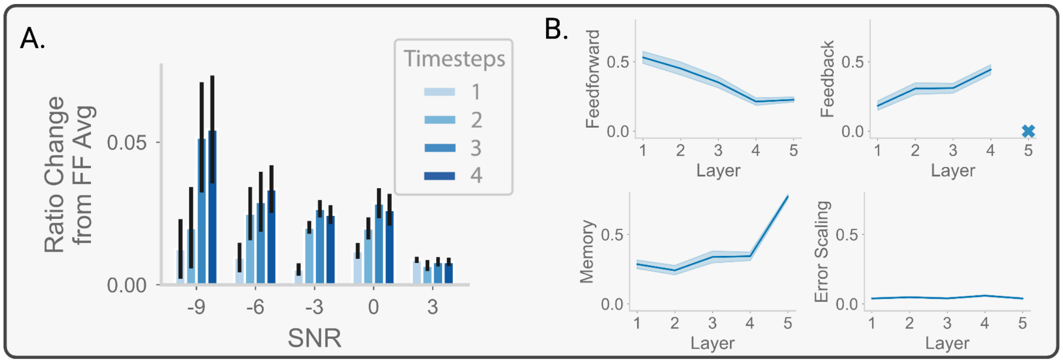

Figure 2: **(A)** Relative increase in performance from the original feedforward on test set for various levels of SNR. 4 timesteps of predictive dynamics were used. SEM is shown for 10 different random initializations of hyperparameter training for each background noise/SNR combination. **(B)** Hyperparameter values of the networks shown in (A). The last (5th) layer of the network does not have feedback by definition.

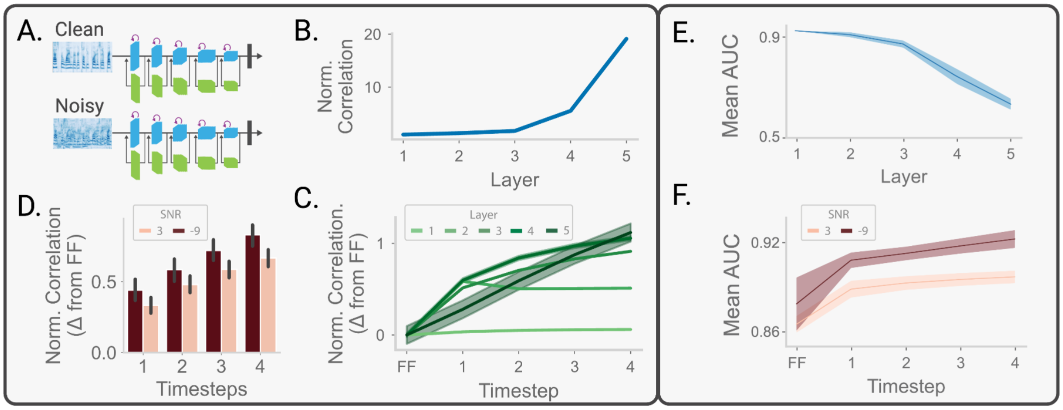

Figure 3: **(A)** Layer representations were compared when the network input was a sound with background noise versus the same sound with no noise. **(B)** Normalized correlation for each layer during the initial feedforward pass. Correlations were normalized at each layer by random shuffles of sound labels. **(C)** Difference in normalized correlation from (B) as predictive time steps are incorporated. **(D)** As in (C), but averaging over all layers and comparing the most noisy (SNR$= -9$) and least noisy (SNR$= 3$) conditions. **(E)** Area under the curve (AUC) of the cumulative singular value spectrum of activity at each layer for feedforward pass alone. **(F)** As in (E), but with predictive timesteps incorporated and comparing the most noisy and least noisy conditions.

3 Results

3.1 Predictive dynamics improve performance on an auditory classification task.

We found that predictive dynamics improved classification performance relative to the feedforward network alone (Fig 2A). This improvement was more dramatic for more corrupted sounds (i.e. lower SNR) and varied across background noise types (see Appendix A.3).

To better understand the contribution of each model component to this performance improvement, we analyzed fitted hyperparameters across layers (Fig 2B). We found that predictive feedback contributed to dynamics for all layers (i.e. feedback weight > 0) and that this contribution increased in deeper layers. Note that a feedforward weight of 1 for all layers is equivalent to the feedforward network. Interestingly, the error correction term (a key element of the canonical predictive coding model [14]) does not play a significant role in network dynamics. In fact, ablating this term had limited impact on the performance of the network (Appendix A.4).

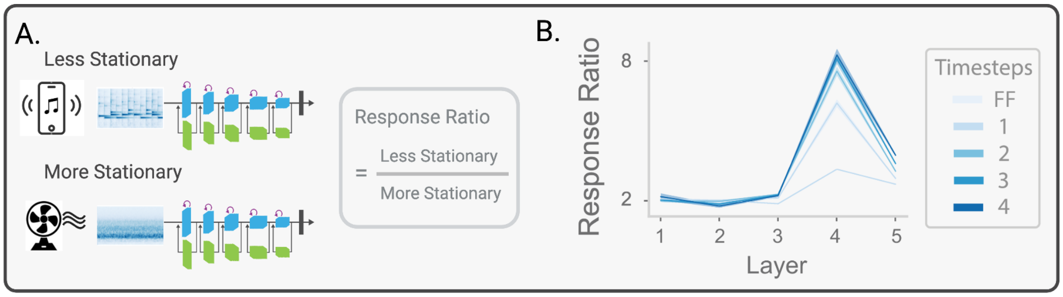

Figure 4: **(A)** Network representations of each layer were compared for less stationary (e.g. ringtone) and more stationary (e.g. air conditioner) sounds. Response ratio was calculated as the mean response to less stationary sounds divided by the response to more stationary sounds. **(B)** Response ratio for each layer of the network, across predictive timesteps.

3.2 Noise robustness is associated with denoising of neural representations and altered dimensionality.

Next, we sought to explore the mechanism driving this performance improvement. We hypothesized that predictive dynamics denoise by pushing network representations towards a manifold of clean sounds learned by the feedback weights [10]. To test this hypothesis, we measured network responses to noisy sounds and their clean counterparts. Consistent with the experimental literature [15], we defined denoising as an increase in correlation between responses to clean and noisy sounds across layers (Fig. 3A).

We found that the feedforward network alone shows denoising across layers (Fig. 3B). However, predictive dynamics still appear useful in this denoising process. Across predictive timesteps, there is an additional, albeit smaller, denoising effect (Fig 3C). Consistent with our performance findings, this effect is stronger for inputs that are noisier (Fig 3D).

We also examined the dimensionality of network activity across each layer, a related measure to denoising. Specifically, we computed the cumulative singular value spectrum of the output at each layer. We summarize the dimensionality by computing the area under the curve of this function (AUC). A low AUC corresponds to high dimensionality. We found that the dimensionality of representations increase across layers in the feedforward network alone (Fig 3E). This is consistent with the observation that increasing the dimensionality of representations is useful for classification performance [16]. However, the introduction of predictive dynamics causes dimensionality to decrease with more predictive timesteps, particularly for noisier conditions (Fig 3F). These findings could be consistent with a mechanism by which the initial feedforward weights are trained to maximize separability of representations, predictive dynamics will contract the representations towards a learned clean manifold.

3.3 Network activity recapitulates neural data in a non-speech task.

A potentially complementary hypothesis suggests that sensory systems denoise by developing statistical summaries of complex scenes [17–19]. Given that background sounds tend to have more stationary statistics while foreground sounds tend to have less stationary statistics, it has been proposed that the auditory cortex could leverage these statistical differences to distinguish them [20]. A recent fMRI study found support for this hypothesis [13]. Specifically, authors found that early layers of the auditory hierarchy respond similarly to stationary and non-stationary sounds while deeper layers have a higher response ratio of non-stationary to stationary sounds.

We test if our model captures this finding by comparing the responses at each layer to a set of natural sounds with varying stationarity. In the feedforward network, we see a modest increase in response ratio across layers of the network. This effect becomes more dramatic with predictive timesteps (Fig 4B). Consistent with neural data [13], these findings suggest that our model is sensitive to the temporal statistics of natural sounds.

4 Discussion

In this work, we sought to evaluate the potential mechanistic role of predictive dynamics in noise robustness using a large scale model of the human auditory system. We found that predictive dynamics improved classification performance across timesteps and that this improvement was most dramatic for more difficult conditions (i.e. lower SNR). The dependence of performance improvement on task difficulty is consistent with other modeling results [10, 12] as well as and empirical findings that that observers rely more on predictions as sensory input becomes less reliable [6]. Additionally, we probed the model by ablation and found that the removal of the error correction term had limited impact on performance. In contrast, the balance of feedforward drive and feedback are essential for performance suggesting more consistency with alternative theories of sensory predictions [21] than canonical predictive coding [14]. This network architecture provides a promising bridge between circuit-level and normative theories of sensory processing.

We probed the mechanism underlying this performance improvement through analyses of network representations. We found evidence that predictive dynamics denoise representations and decrease dimensionality. These findings are consistent with a mechanism by which predictive dynamics denoise by pushing network representations towards a manifold of clean sounds learned by the feedback weights [10]. In future work, we aim to more rigorously evaluate this mechanism through manipulations of the manifold learned by feedback weights.

Finally, we evaluated the extent to which this modeling framework captures findings in literature beyond speech tasks. Specifically, we evaluated model responses to stationary (i.e. background) and non-stationary (i.e. foreground) natural sounds. Consistent with brain data [13, 22, 23], we found that later layers of the model responded preferentially to non-stationary sounds and that this effect became more dramatic across predictive timesteps. Intriguingly, we recapitulate these results despite the fact that the network was only trained on speech and was not exposed during training to other types of natural sounds. This suggests that background invariance could be a general consequence of predictive dynamics.

We acknowledge several limitations on the interpretation of our work. First, training occurred in a step-wise fashion with weights being frozen after each step. While this approach is common in machine learning, it is possible that end-to-end training would yield different solutions. Second, in this predictive coding scheme, uncertainty is not taken into account in the updating process. In Bayes-optimal predictive coding, the integration of sensory evidence with top-down priors is weighted by their respective uncertainty. Allowing for this weighting based for individual stimuli could also yield a different solution. Finally, while we have shown that our model qualitatively captures some findings in the literature, quantitative comparisons between model and human data are needed to demonstrate that this model can provide insight into the human auditory system.

5 Broader Implications

Alterations in predictive sensory processing have been widely documented in disease states, such as hallucinations [24–26]. While modeling approaches have provided insight into the cognitive mechanisms that may be disrupted in these states [27, 28], alterations in cortical processing are unclear. Our network model provides a way to probe how circuit-level alterations in predictive dynamics may give rise to pathological states.

We further propose that studies of auditory processing can provide insight into general sensory processing principles. In vision, predictive dynamics are often considered in the context of static images [11, 10, 12]. However, many visual problems require information integration over time. For instance, recognition of a well-camouflaged predator likely requires temporal integration of visual (e.g. motion of predator relative to surroundings). Insights into how the human visual system solves this problem could come from findings in audition, an inherently temporal problem.

References

[1] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):4, 2009.

[2] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[3] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[4] Helen Blank and Matthew H Davis. Prediction errors but not sharpened signals simulate multivoxel fmri patterns during speech perception. *PLoS biology*, 14(11):e1002577, 2016.

[5] Jan Drugowitsch and Alexandre Pouget. Probabilistic vs. non-probabilistic approaches to the neurobiology of perceptual decision-making. *Current opinion in neurobiology*, 22(6):963–969, 2012.

[6] Floris P De Lange, Micha Heilbron, and Peter Kok. How do expectations shape perception? *Trends in cognitive sciences*, 22(9):764–779, 2018.

[7] Jon H Kaas and Troy A Hackett. Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences*, 97(22):11793–11799, 2000.

[8] Ruben S van Bergen and Nikolaus Kriegeskorte. Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65:176–193, 2020.

[9] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

[10] Bhavin Choksi, Milad Mozafari, Callum Biggs O'May, Benjamin Ador, Andrea Alamia, and Rufin VanRullen. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. *arXiv*, 2021.

[11] Grace W. Lindsay, Thomas D. Mrsic-Flogel, and Maneesh Sahani. Bio-inspired neural networks implement different recurrent visual processing strategies than task-trained ones do. *bioRxiv*, page 2022.03.07.483196, 2022. doi: 10.1101/2022.03.07.483196.

[12] Andrea Alamia, Milad Mozafari, Bhavin Choksi, and Rufin VanRullen. On the role of feedback in visual processing: a predictive coding perspective. *arXiv preprint arXiv:2106.04225*, 2021.

[13] Alexander J E Kell, Daniel L K Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.03.044.

[14] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. ISSN 1097-6256. doi: 10.1038/4580.

[15] Alexander JE Kell and Josh H McDermott. Invariance to background noise as a signature of non-primary auditory cortex. *Nature communications*, 10(1):1–11, 2019.

[16] Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, 37:66–74, 2016.

[17] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493–498, 2013.

[18] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

[19] M.W. Spratling. A review of predictive coding algorithms. *Brain and Cognition*, 112:92–97, 2017. ISSN 0278-2626. doi: 10.1016/j.bandc.2015.11.003.

[20] Richard McWalter and Josh H McDermott. Adaptive and selective time averaging of auditory scenes. *Current Biology*, 28(9):1405–1418, 2018.

[21] David J Heeger. Theory of cortical function. *Proceedings of the National Academy of Sciences*, 114(8):1773–1782, 2017.

[22] Alexander J. E. Kell and Josh H. McDermott. Invariance to background noise as a signature of non-primary auditory cortex. *Nature Communications*, 10(1):3958, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11710-y.

[23] Sam V. Norman-Haignere, Laura K. Long, Orrin Devinsky, Werner Doyle, Ifeoma Irobunda, Edward M. Merricks, Neil A. Feldstein, Guy M. McKhann, Catherine A. Schevon, Adeen Flinker, and Nima Mesgarani. Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, 6(3):455–469, 2022. ISSN 2397-3374. doi: 10.1038/s41562-021-01261-y.

[24] Guillermo Horga, Kelly C. Schatz, Anissa Abi-Dargham, and Bradley S. Peterson. Deficits in Predictive Coding Underlie Hallucinations in Schizophrenia. *The Journal of Neuroscience*, 34 (24):8072–8082, 2014. ISSN 0270-6474. doi: 10.1523/jneurosci.0200-14.2014.

[25] Angeliki Zarkali, Rick A Adams, Stamatios Psarras, Louise-Ann Leyland, Geraint Rees, and Rimona S Weil. Increased weighting on prior knowledge in lewy body-associated visual hallucinations. *Brain communications*, 1(1):fcz007, 2019.

[26] Clifford M Cassidy, Peter D Balsam, Jodi J Weinstein, Rachel J Rosengard, Mark Slifstein, Nathaniel D Daw, Anissa Abi-Dargham, and Guillermo Horga. A perceptual inference mechanism for hallucinations linked to striatal dopamine. *Current Biology*, 28(4):503–514, 2018.

[27] Albert R Powers, Christoph Mathys, and Philip Robert Corlett. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351):596–600, 2017.

[28] Guillermo Horga and Anissa Abi-Dargham. An integrative framework for perceptual disturbances in psychosis. *Nature Reviews Neuroscience*, 20(12):763–778, 2019. ISSN 1471-003X. doi: 10.1038/s41583-019-0234-1.

[29] Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. URL `https://aclanthology.org/H92-1073`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] To maintain anonymity.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We didn't properly measure this.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 2
   (b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Predictive dynamics are integrated into feedforward activity over many timesteps

We use the recently introduced Predify framework [10, 12]. A schematic is shown in Fig 5. Under this framework, the network receives input at time $t = 0$. The activity e of layer i at time t is defined as:

$$e_i(t+1) = \gamma f_i + \beta d_{i+1} - \alpha \nabla \epsilon_i + \mu e_i(t)$$

The feedforward term $f_i = \mathcal{F}_i(e_{i-1}(t+1); W_i^f)$ is the output of a convolutional layer $\mathcal{F}$ parameterized by weights W_i^f. The feedback term $d_{i+1} = \mathcal{D}_{i+1}(e_{i+1}(t); W_{i+1}^b)$ is the output of a deconvolutional layer $\mathcal{D}$ parameterized by weights W_i^b. Finally, the error correction term $\nabla \epsilon_i = \nabla_{e_i} MSE(d_i, e_{i-1})$. Each term is weighted by hyperparameters $\gamma, \beta, \alpha, \mu$. Importantly, there is the constraint $\gamma + \beta + \mu = 1$.

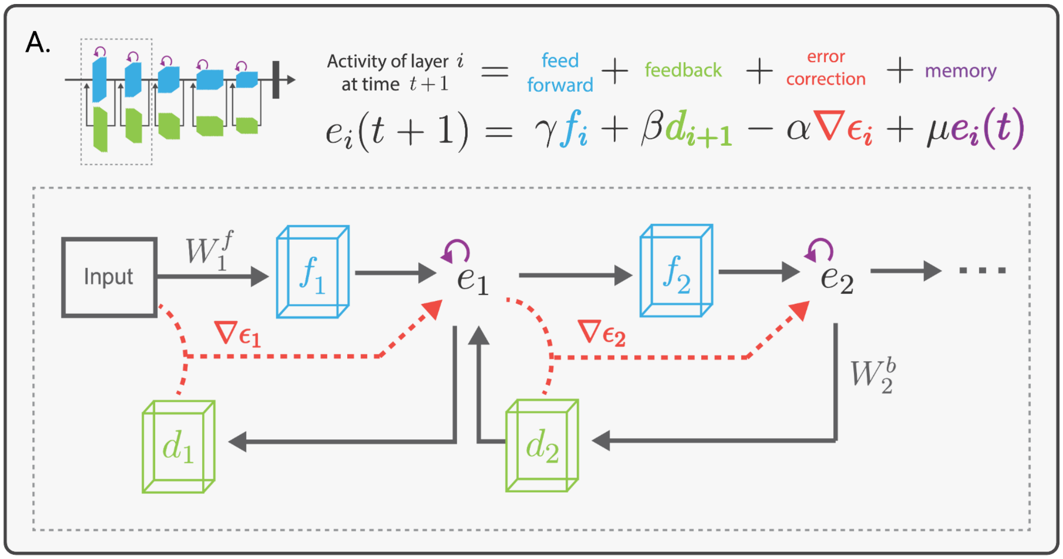

Figure 5: **(A)** Schematic of the first two layers of the network and the predictive dynamics equation. The activity of layer i at time $t+1$ is denoted as $e_i(t+1)$ and is computed as a linear combination of feedforward, feedback, error correction, and memory terms. In the schematic, blue blocks represent convolution layers (parameterized by weights W_i^f for layer i), while green blocks represent deconvolution layers (parameterized by weights W_i^b for layer i).

A.2 Training details for predictive deep network model of auditory processing

Here we provide more details of the training for the predictive network (Fig 6). Speech for the word classification task was sampled from the Wall Street Journal corpus [29]. The sounds are converted to cochleagrams as in [13] using pyCocleagram (`https://github.com/mcdermottLab/pycochleagram`). Cochleagrams are time-frequency decomposition of sounds with frequency dependent bandwiths and compressive non-linearities similar to the human cochlea.

Background noise is added as in [13]. Additionally, we introduce pink noise as a new background noise type to establish a particularly easy noise condition with highly stationary statistics. The feedforward network is trained for word classification only on noisy sounds [13]. Feedback weights are trained on an unsupervised reconstruction task with clean sounds [10]. Finally, hyperparameters are learned via backpropagation to maximize performance on word classification on noisy sounds [12]. Importantly, hyperparameters are trained separately for each combination of background noise and SNR level [12]. All analyses are performed on a held-out validation set not seen in any of the training procedures. All error bars and error shading shown are standard error.

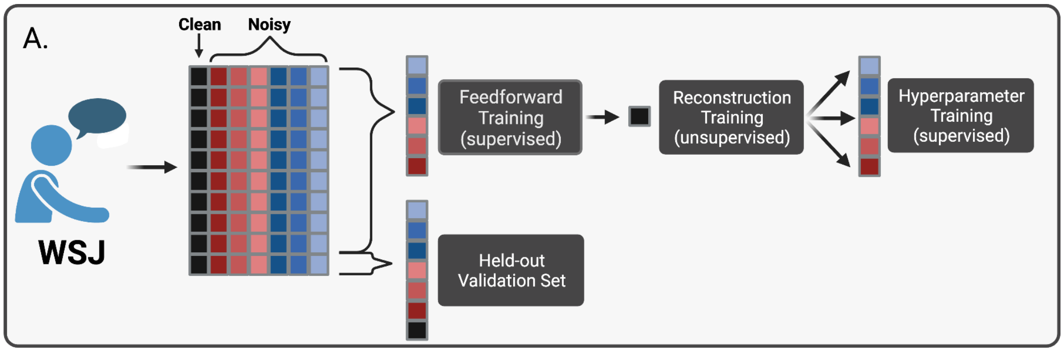
A.
Clean
Noisy
WSJ
Feedforward Training (supervised)
Reconstruction Training (unsupervised)
Hyperparameter Training (supervised)
Held-out Validation Set

Figure 6: **(A)** Schematic of dataset generation and how it is split amongst the different training procedures. Each square represents some 2-second sound from the WSJ corpus. The color represents the SNR of the sound. Sounds either have no added noise (black) or have added noise of SNR $-9, -6, -3, 0, 3$ decibels (red through blue).

A.3 Model performance, split across each background type

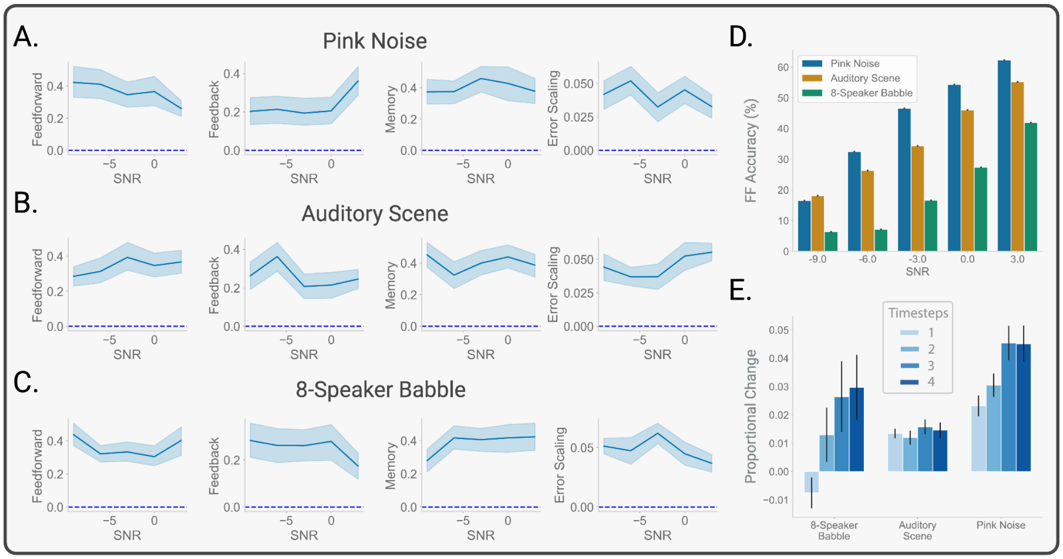

Figure 7: **(A)** Hyperparameter values of the trained networks, but only for the networks trained on pink noise. **(B)** As in (A), but only for auditory scene. **(C)** As in (A), but only for 8-speaker babble. **(D)** Feedforward accuracy on held-out data set, split by background noise. **(E)** Ratio change of accuracy from feedforward across predictive timesteps, evaluated on held-out data set and separated by background type.

A.4 Model performance with error correction term ablated

To test how crucial the error ablation term is, we set the associated hyperparameter α to 0 during the hyperparameter training and during the network evaluation. We find that the hyperparameter values and performance of the network do not appear to be impacted by this ablation (Fig 8).

A.5 Potential societal impacts

We do not foresee negative societal impacts from this study, as we are focused on basic neuroscience contributions.

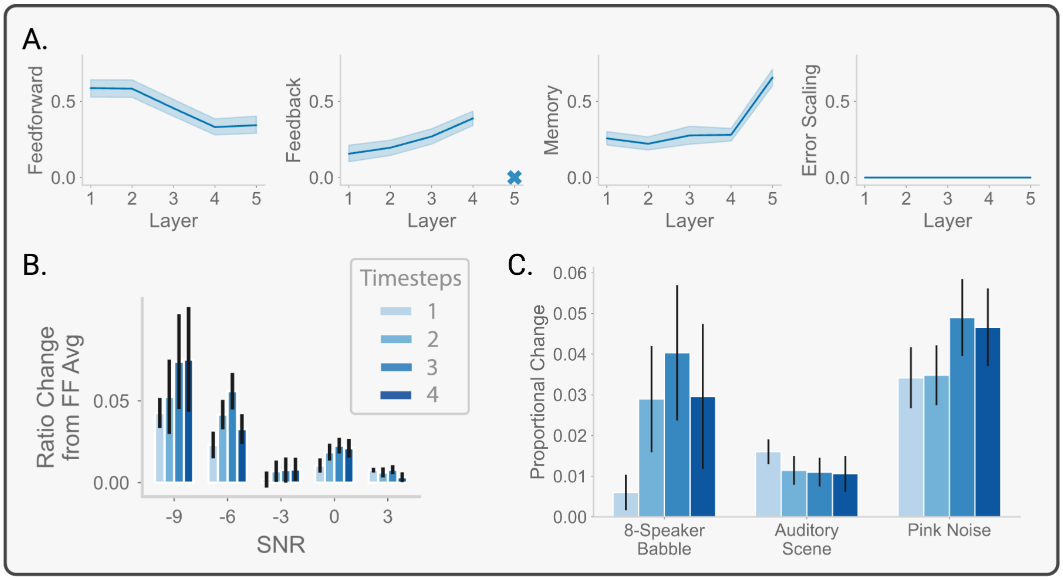

Figure 8: **(A)** As in Fig. 2B but with error correction term ablated. **(B)** As in Fig. 2A, but with error correction term ablated. **(C)** As in Fig. 7E, but with error correction term ablated.