Modeling neural coding in the auditory midbrain with high resolution and accuracy

Fotios Drakopoulos[1]*, Shievanie Sabesan[1], Yiqing Xia[1], Andreas Fragner[2], Nicholas A Lesica[1]

[1]Ear Institute, University College London, United Kingdom, [2]Perceptual Technologies, United Kingdom

*Correspondence: f.drakopoulos@ucl.ac.uk

Abstract

Computational models of auditory processing can be valuable tools for research and technology development. Models of the cochlea are highly accurate and widely used, but models of the auditory brain lag far behind in both performance and penetration. Here, we present ICNet, a model that provides accurate simulation of neural dynamics in the inferior colliculus across a wide range of sounds, including near-perfect simulation of responses to speech. We developed ICNet using deep learning and large-scale intracranial recordings from gerbils, addressing three key modeling challenges that are common across all sensory systems: capturing the full statistical complexity of neuronal response patterns; accounting for physiological and experimental non-stationarity; and extracting features of sensory processing that are common across different brains. ICNet can be used to simulate activity from thousands of neural units or to provide a compact representation of central auditory processing through its latent dynamics, facilitating a wide range of hearing and audio applications.

Introduction

Computational models of sensory systems are widely used in fundamental research. They provide a platform for synthesizing existing knowledge and allow for generation and initial testing of hypotheses without the need for in vivo experiments. They can also facilitate the development of technologies that seek to approximate natural sensory function in artificial systems. But the utility of a model for a particular application depends on the degree to which it provides an accurate simulation of the true biological system at the relevant spatial and temporal scales. Numerous high-resolution models of the cochlea have been developed that faithfully capture its mechanical and electrical properties [1], with widespread uptake across academia and industry. Models of auditory processing in the brain, on the other hand, are much less accurate, with even the best missing out on a significant fraction of the explainable variance in sub-cortical and cortical neural activity [2, 3].

Models of the cochlea have typically been designed by hand to mimic the transformations performed by the biological system. This approach works well for the cochlea because it is relatively simple, containing only a few stages with homogeneous elements. The brain, however, with its many interconnected and highly non-linear neural networks, is too complex for such an approach. Hand-designed models of the auditory brain can be effective in describing specific features of responses to a limited class of sounds [4–7], but more general models of the brain require the use of “black box” frameworks with parameters estimated from neural recordings. To facilitate model fitting from limited data, the allowable transformations are typically constrained to be either linear or low-order non-linear [2]. These constraints aid interpretability, but such

models lack the capacity to capture the full complexity of auditory processing and, thus, are not accurate enough for many applications.

Advances in deep learning have created a new opportunity to develop accurate models of sensory processing in the brain, making it feasible to fit models that capture arbitrary non-linear transformations directly from data. Recent studies have shown that deep neural network (DNN) models of the sensory periphery can be highly accurate. DNNs perform similarly to biophysical models of the cochlea in predicting a wide range of auditory phenomena [8–11] and, beyond the auditory system, models of the retina provide accurate predictions of ganglion cell responses to natural scenes [12]. Initial attempts to build DNN models of sensory processing in the brain have also produced impressive results. Models of primary visual cortex (V1) responses to natural images explained 50-90% of the explainable variance in recorded activity [13–15], while models of V4 explained 90% [16]. DNN models of primary auditory cortex (A1) responses to natural sounds perform similarly well, explaining 60-70% of the explainable variance in recorded activity [17, 18]. But these models reproduce only low-resolution measurements of activity (calcium transients [13, 15], spike counts over large time windows [14, 16], field potentials [18] or fMRI voxels [17]), ignoring potentially important information in the neural code that is present at fine spatial and temporal scales. One recent attempt to use DNNs to model single- and multi-unit spiking in auditory cortex with a temporal resolution of 10 ms succeeded in capturing approximately half of the explainable variance [19].

We recently developed a DNN model of the inferior colliculus (IC), the midbrain hub of the central auditory pathway, to simulate neural response patterns with millisecond precision [20]. We trained the model on multi-unit activity (MUA) recorded across the extent of the IC, with the goal of identifying and capturing latent neural dynamics that reflect the fundamental computations that are common to all brains, rather than the idiosyncrasies of particular neurons. This model was effective in predicting mean responses to a limited set of speech sounds. However, to bring its performance up to the level required for it to serve as a useful general model of central auditory processing or as a valid in silico surrogate for the IC in auditory neuroscience, there are three key challenges that must be addressed. The first is to capture the full statistical complexity of neural activity. Models of neural coding typically rely on regression frameworks with assumed distributions (e.g., Poisson) that constrain the relationship between response strength and reliability [2]. While these assumed distributions might be valid for some higher-level brain areas, they are a poor match for early sensory pathways where strength and reliability exhibit complex interactions.

The second challenge is to address non-stationarity. The mapping between sensory input and recorded neural activity can change over time for both physiological reasons (e.g., changes in brain state) and experimental reasons (e.g., movement of the recording electrode within the brain). Because training high-capacity models requires the use of large datasets recorded over long periods of time, non-stationarity is inevitable and must be explicitly accounted for. The final challenge is to overcome the idiosyncrasy of individual brains and the recordings made from them. While models must be trained on individual recordings, the final model of a sensory brain area should reflect only the processing that is common across all healthy individuals.

To address these challenges and derive an accurate and general model of central auditory processing, we developed a novel encoder-decoder framework for modeling neural coding. The framework comprises a shared encoder that maps a sound waveform to a generic latent representation, followed by separate decoders that map the generic representation to the neural activity of individual brains. Each decoder also receives a second input, specific to each brain, that it uses to modulate the generic representation to account for non-stationarity. The final model, which we call ICNet, can simulate neural activity across many brains in response to a wide range of sounds with high accuracy and fast execution, and its latent dynamics provide a compact description of auditory processing that can serve a variety of applications.

In the following sections, we describe the DNN-based framework that we used to develop ICNet and then evaluate its performance. We first describe the DNN components that we designed to address each of the three modeling challenges. We then present the final model and demonstrate how its encoder can be

used as a front-end for automatic speech recognition. Finally, we establish the remarkable accuracy and generality of the model by comparing neural responses simulated by ICNet with those recorded from the IC for a variety of sounds that were not part of training (including speech in quiet and in noise, moving ripples, and music at different intensities).

Results

To collect the dataset required for model training, we recorded neural activity from the IC of nine normal-hearing gerbils, a common animal model for human hearing, using custom-designed electrode arrays with 512 channels (Fig 1a) [20, 21]. We made the recordings under anesthesia, allowing us to present a wide variety of sounds to each animal across a range of intensities, including speech in quiet and in noise, processed speech, music, environmental sounds, moving ripples and tones that together totaled more than 10 hr in duration. We processed the recordings to extract MUA (threshold crossing counts on each recording channel in 1.3 ms time bins) [20, 22]. We included all recorded units in model training, but only those units with a signal correlation (correlation of responses across repeated trials of broadband noise) of 0.2 or higher are included in our model evaluation (494 $\pm$ 16.7 units from each animal, resulting in a total of 4446 units across all animals).

We then trained DNN models to simulate neural coding, i.e., to take as input each presented sound waveform and produce as output the corresponding MUA spiking patterns recorded from all units in each individual animal. These models use a cascade of convolutional layers to encode the sound waveform s to a latent representation $\hat{r}_b$ (bottleneck), which is then passed to a simple linear decoder to predict the response $\hat{R}$ for the recorded units (Fig 1b). The output of the decoder for each unit in each time bin defines a probability distribution $p(\hat{R}|s)$, from which we can then sample to simulate the activity $\hat{R}$ of each unit across time with millisecond precision. Once trained, the DNN models can be used to simulate the neural activity of all recorded units for any sound, whether it was presented during neural recordings or not.

DNNs can capture the statistics of neural response patterns

We first trained a set of baseline models to simulate activity across all units for each animal through Poisson regression [20] (N_c = 1 in the decoder layer of Fig 1b to produce the Poisson rate parameter λ). The Poisson assumption is common in models of neural coding. It provides a theoretical basis for the optimal estimation of model parameters via classical statistical methods, and it is a reasonable approximation of the true distribution of activity in some brain areas. But activity in sub-cortical areas like the IC is typically underdispersed, i.e., it is much more reliable across trials than is expected for a Poisson process [23, 24]. This phenomenon is illustrated in Fig 1c, which compares the peri-stimulus time histograms (PSTHs) of recorded and simulated neural responses for an example IC unit to 128 trials of a pure tone. The Poisson model captured the mean activity across trials well (indicated by the solid lines), but overestimated the variability across trials (indicated by the shaded areas).

Fortunately, DNNs can be trained to predict full activity distributions without any assumptions by framing the problem as classification rather than regression [25]. We trained the same DNN architecture for each animal with a crossentropy loss and categorical distributions with N_c = 5 discrete classes (corresponding to the probability of each unit having a count of 0 to 4 within a given 1.3 ms time bin). The crossentropy model accurately captured both the mean activity and the variability across trials for the example unit (Fig 1c). For all time bins within the duration of the tone, the Fano factor (ratio of variance to mean) was 0.55 for the true response, 0.54 for crossentropy model, and 1.16 for the Poisson model. (Note that because the rate of the Poisson model varies across time bins, the overall Fano factor of its activity does not need to be exactly 1.) For all units from one example animal in response to four different tone frequencies, the Fano factor for the

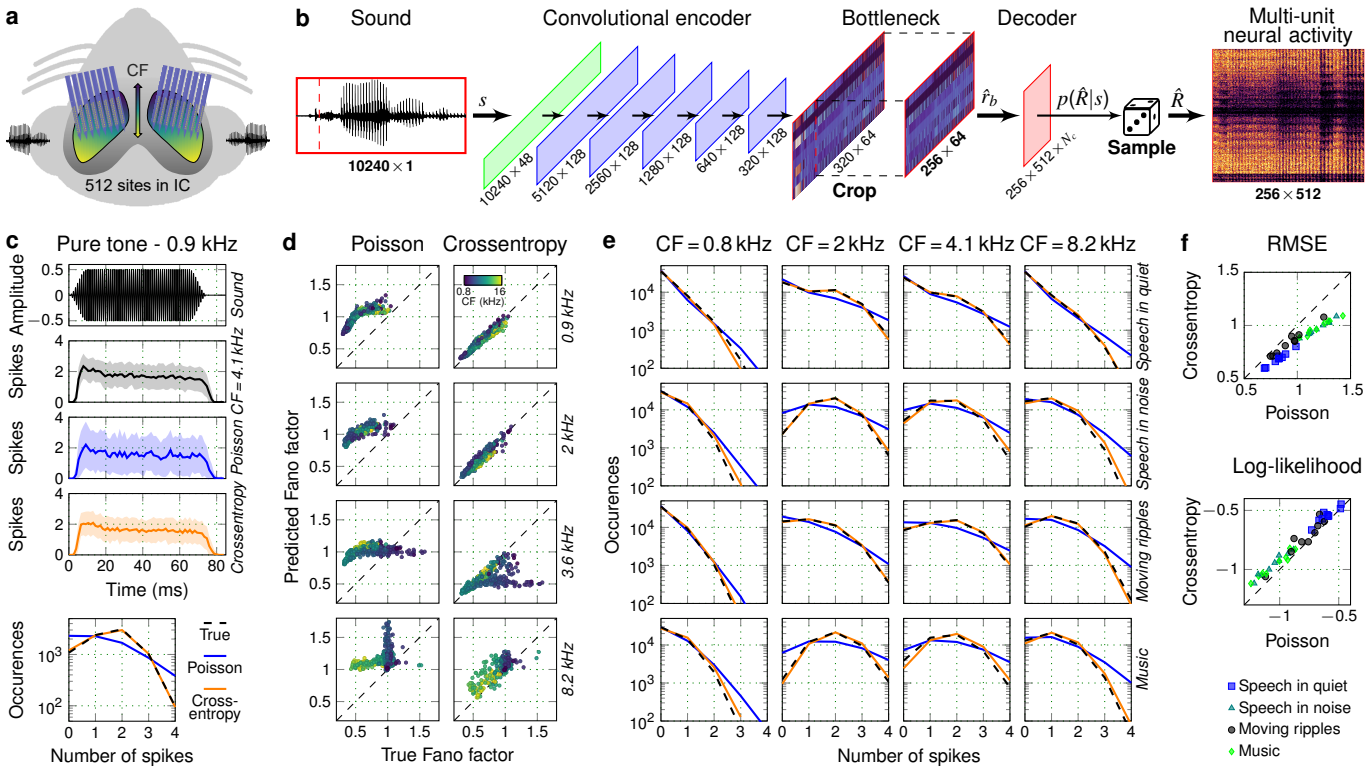

Fig 1. Modeling the statistical properties of neural activity. a. Schematic of the geometry of custom-designed electrode arrays for large-scale recordings in the gerbil inferior colliculus (IC). The recording sites were distributed across 8 electrode shanks and spanned a plane measuring 1.4 × 0.45 mm within each hemisphere. **b.** Schematic diagram of the deep neural network (DNN) architecture trained to map a sound waveform to neural activity for 512 units from the IC of a single animal. N_c is the number of parameters in the probability distribution $p(\hat{R}|s)$, which is 1 for Poisson models and 5 for crossentropy models. **c.** Peri-stimulus time histograms (PSTHs) of recorded and predicted neural activity for one unit (characteristic frequency (CF) of 4.1 kHz) in response to 128 presentations of a 0.9 kHz pure tone at 85 dB SPL. Neural activity was simulated using either a Poisson or crossentropy model. The shaded regions indicate the standard deviation across the 128 trials. The bottom panel shows the distribution of MUA counts across trials for the responses in the PSTH plots, averaged across time bins during the presentation of the tone. **d.** Comparison of the Fano factor (variance-to-mean ratio) between recorded and predicted neural activity (each point corresponds to one of 483 units from one animal) for the Poisson and crossentropy models (columns) in response to 4 tones (rows). Points are colored based on the CF of the corresponding unit, with brighter colors indicating higher CFs. **e.** Comparison of MUA count distributions for recorded and predicted neural activity for 4 units (columns) in response to 4 sounds (rows). **f.** Performance comparison across 9 animals and 4 sounds between Poisson and crossentropy models. Each symbol represents the average across all time bins and units for one animal in response to one sound.

crossentropy model was much closer to the true value than the Fano factor for the Poisson model (Fig 1d; the average absolute error was 0.38 for the Poisson model and 0.07 for the crossentropy model).

To compare the overall performance of the Poisson and crossentropy models, we simulated responses for all animals to four different sounds: speech in quiet, speech in noise, moving ripples, and music. As shown in Fig 1e for four example units, the crossentropy model accurately predicted the full MUA count distributions for each sound. We compared the overall performance of the models for each animal and sound using two metrics (averaged across all time bins and units): the root-mean-square error (RMSE) between the recorded and predicted activity, and the logarithm of the probability (log-likelihood) of observing the recorded activity from the activity distribution inferred by the model. We found consistent improvements for the crossentropy models across all animals and sounds (Fig 1f; improvement of 15.98% [15.88, 16.02] for RMSE (median and 95% CIs over bootstrap samples) and 8.13% [8.04, 8.16] for log-likelihood). These results demonstrate that characterizing the full distributions of spiking activity for each neural unit is feasible

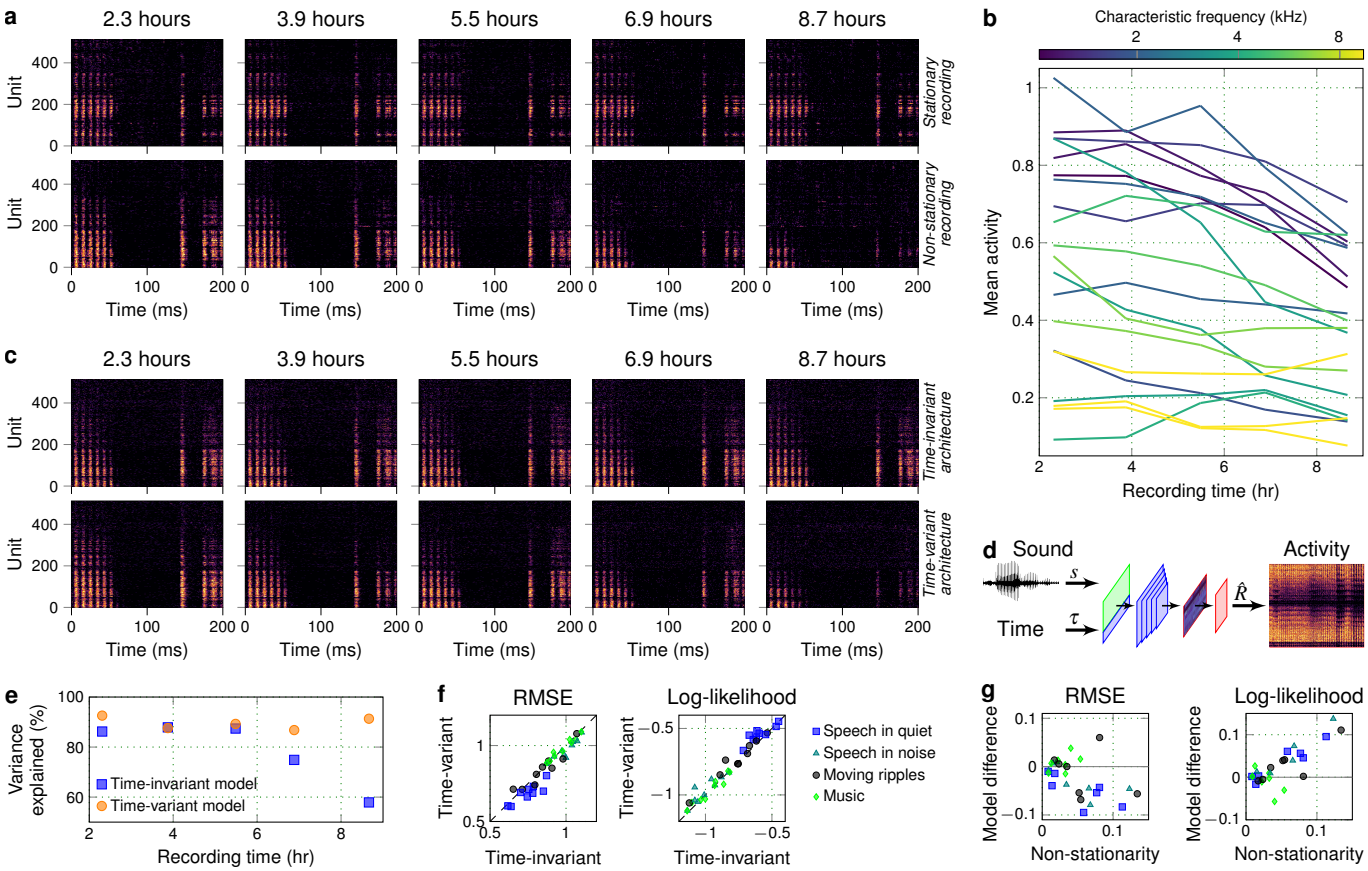

Fig 2. Accounting for non-stationarity in neural recordings. a. Example neural activity in response to the same sound (speech at 60 dB SPL) presented at five different times within two individual recordings. Units were ordered based on their CFs. Brighter colors indicate higher activity. **b.** The mean activity (per bin) of 20 randomly selected units from the non-stationary recording shown in **a**. **c.** Predicted neural activity for the non-stationary recording in **a** using the time-invariant and the time-variant models. **d.** Schematic diagram of the time-variant DNN architecture. **e.** Median predictive power of the time-invariant and time-variant models across all units at five different times during the non-stationary recording shown in **a**. Predictive power was computed from the simulated and recorded neural responses as the fraction of explainable variance explained for a 30-s speech segment presented at 60 dB SPL. **f.** Performance comparison across 9 animals and 4 sounds between time-variant and time-invariant DNN models. Each symbol represents the average across all time bins and units for one animal in response to one sound. **g.** Performance difference between models (time-variant - time-invariant) as a function of recording non-stationarity across 7 animals and 4 sounds.

and improves the accuracy of simulated neural coding in the IC.

DNNs can account for non-stationarity in neural recordings

Capturing the full non-linearity of auditory processing requires the use of high-capacity models, which in turn requires large datasets recorded over long periods of time. Non-stationarity in such recordings is inevitable, as the physiological state of the animal or the electrode position can vary over the course of the recording, altering the mapping from sound to neural activity expressed by individual units. Fig 2a illustrates this for two of our recordings, where neural activity in response to the same speech sound was either stable (stationary) or variable (non-stationary) over the 10 hours of recording. While there is an overall trend of decreased activity in the second example, the non-stationarity can vary widely across individual units (Fig 2b), creating a challenging modeling problem.

A standard encoder-decoder model can only capture a time-invariant encoding that will reflect the

average mapping from sound to neural activity over a large dataset. To overcome this limitation, we designed a time-variant version of the crossentropy model that takes an additional input which reflects the time during recording at which the sound was presented (Fig 2d). The interactions between the two inputs within the encoder enable the model to alter its sound encoding as needed to account for non-stationarity. After the model is trained, neural activity can be simulated to reflect the sound encoding at specific time points during the recording (by providing the appropriate additional inputs), or the simulation can made stationary by keeping the additional input constant across all sounds.

As a first test, we trained time-invariant and time-variant models to predict the neural activity of the example non-stationary recording and evaluated their performance in predicting the responses shown in Fig 2a. The time-invariant model, by construction, predicts the same responses each time the sound is presented. But the time-variant model learned to utilize the additional input to capture the changes over the course of the recording, predicting strong responses at early recording times and weak responses at late recording times (Fig 2c). We quantified the differences in model performance for this example by computing the fraction of explainable variance in the recorded activity that each model explained across all units for each of the five speech presentations. The results demonstrate that the time-variant model can capture the changes in the mapping from sound to neural activity over time, consistently explaining close to 90% of the explainable variance across the entire recording (Fig 2e), while the performance of the time-invariant model was much more variable and worse overall.

Across all animals and sounds, we found that the time-variant models improved performance by an average of 2.67% [2.60, 2.69] for RMSE and by 3.22% [3.11, 3.26] for log-likelihood (Fig 2f). But we did not expect performance improvements across all animals, as the time-variant model should only achieve better predictions for recordings that are non-stationary. We categorized the recordings for which we had repeated trials of the same sounds presented at different times in the recording (seven out of nine animals) based on their non-stationarity. We measured non-stationarity as the absolute difference of the covariance of neural activity between two successive trials and two trials recorded several hours apart. As expected, we found that the benefit of the time-variant model increased with non-stationarity for both the RMSE and the log-likelihood measures (Fig 2f), approaching 10% improvement on RMSE and 15% improvement on log-likelihood for the most non-stationary recordings. These results demonstrate that time-variant DNNs can flexibly model non-stationarity, improving performance for recordings that are unstable without compromising performance for stable recordings.

Multi-branch DNNs can capture shared latent dynamics across animals

Our goal was not to model any individual animal, but rather to develop a generic model of the IC that describes the features of sound encoding that are common across all animals. Recordings from any individual animal will inevitably contain idiosyncrasies that reflect variations in anatomy or physiology, or in the position of the recording electrode within the IC. These idiosyncrasies may be critical to preserve when developing models for applications such as hearing aid development, which requires sound processing to be unique to each individual hearing loss profile. But for applications that require a generic model of normal auditory processing, architectures that can extract what is common across recordings from different individuals are more suitable.

We have previously demonstrated that the acoustic information in IC activity is restricted to a low-dimensional subspace which is shared across animals with the same hearing status [20]. This result suggests that it should be possible for a DNN with shared latent dynamics to predict neural activity for any normal-hearing animal. We thus designed a multi-branch DNN architecture comprising a shared encoder and separate decoder branches that are used to simulate neural activity for each of nine individual animals (Fig 3a). To simulate activity for all animals with a shared encoder, we also had to modify the way in which the additional time input was used so that the non-stationarity of each individual recording could be accounted

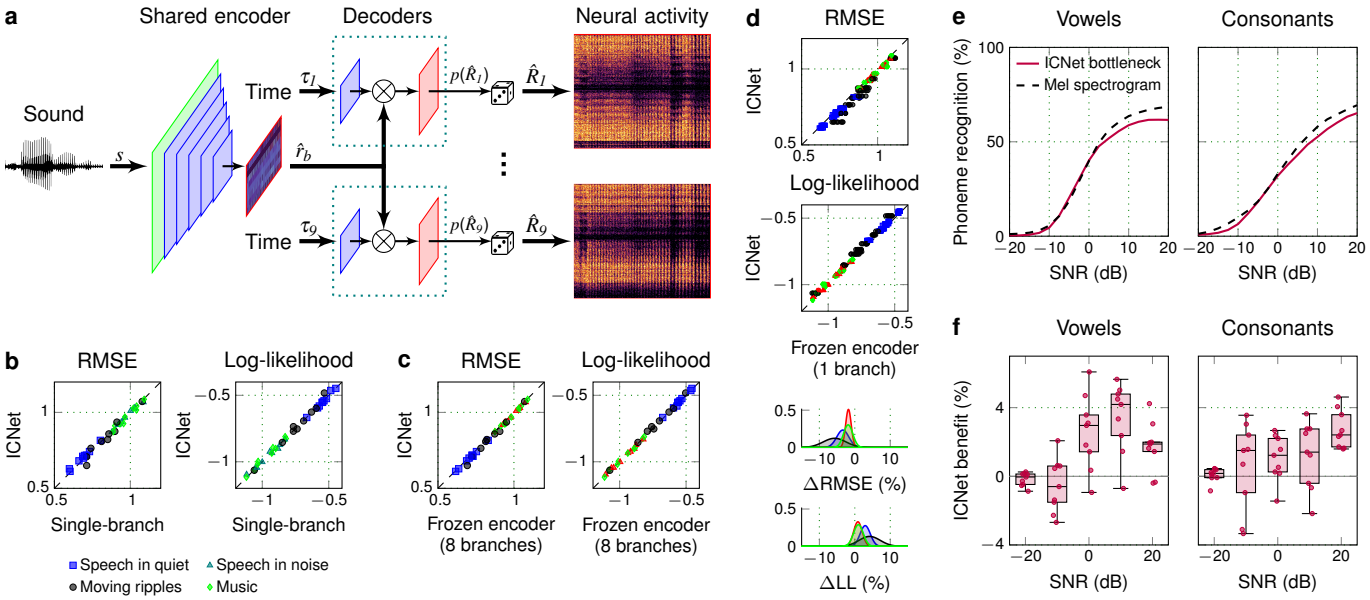

Fig 3. Capturing shared sound encoding across animals. a. Schematic diagram of the ICNet model trained to map a sound waveform to the neural activity of 9 animals. **b.** Performance comparison for 9 animals between ICNet and the corresponding single-branch models. Each symbol represents the average across all time bins and units for one animal in response to one sound. **c.** Cross-validation comparison between ICNet and individual eight-branch models with frozen encoders and decoders trained to predict the ninth animal. The similar performance across animals and sounds demonstrates that multi-branch models learn generic latent dynamics that can be transferred to other animals. **d.** Cross-validation comparison between ICNet and individual single-branch models with frozen encoders and decoders trained to predict the remaining eight animals. Distributions in the bottom show the spread of ICNet benefit for each sound compared to all 72 frozen-encoder models. The decrease in performance across animals and sounds demonstrates that single-branch models cannot fully describe the sound encoding of other animals. **e.** Vowel and consonant recognition performance of an automatic-speech-recognition (ASR) model trained to predict phonemes from either the ICNet bottleneck response $\hat{r}_b$ or a Mel spectrogram. Phoneme recognition was computed for sentences presented at 65 dB SPL and mixed with speech-shaped noise at SNRs between -20 and 20 dB. **f.** Benefit in recognition performance for the ASR model trained with the ICNet bottleneck response as input, compared to ASR models trained with the bottleneck responses of the 9 single-branch models.

for. We introduced animal-specific time inputs at the level of the decoder and combined each time input with the bottleneck output, allowing for a unique time-variant mapping from the generic latent representation to the neural activity for each animal. After training, the multi-branch model, which we call ICNet, can simulate neural activity for nine individual animals in response to any sound in a manner that reflects the state of each animal at any given time.

We first evaluated ICNet by simulating neural activity in response to four sounds (with the time input matched to the time at which each sound was presented to each animal). ICNet matched the performance of the corresponding single-branch models for all nine animals (Fig 3b; average RMSE increase of 0.04% [0.01, 0.06] and log-likelihood decrease of 0.39% [0.34, 0.41] compared to time-variant models trained to predict the activity of a single animal). To establish that ICNet encoded neural dynamics that were shared across all animals (rather than the idiosyncrasies of each), we performed a cross-validation experiment. We held out each of the nine individual animals in turn while training an eight-branch model to predict the neural activity of the remaining animals. We then froze the encoder of each eight-branch model and trained a new decoder that used the output of the frozen encoder to predict the neural activity of the held-out animal. As shown in Fig 3c, the models that used a frozen encoder and an optimized decoder performed similarly to the original ICNet (1.19% [1.16, 1.20] decrease of RMSE and 0.41% [0.38, 0.42] increase of log-likelihood for ICNet compared to the frozen-encoder models), which suggests that ICNet performs a generic encoding

that can predict the neural activity of any normal-hearing animal.

We then reversed the cross-validation experiment to establish the benefit of our multi-branch DNN architecture, i.e., the degree to which the single-branch models captured the idiosyncrasies of each animal without fully capturing the neural dynamics that are shared across animals. We froze the encoder of each of the nine single-branch models and trained eight new decoders to predict the neural activity of the remaining animals. As shown in Fig 3d, the difference in the performance of ICNet and these frozen-encoder single-branch models was small, but the frozen-encoder models were almost always worse, especially for speech in quiet and moving ripples (average RMSE decrease of 3.07% [3.05, 3.07] and log-likelihood increase of 2.01% [2.00, 2.02] for ICNet compared to the frozen-encoder models). Taken together, the results of these cross-validation tests establish that ICNet is a generic yet comprehensive model of sound encoding.

The design of ICNet with its complex encoder and simple decoders differs from that of standard encoder-decoder DNN models. Typically, a complex encoder is used to map the input to an abstract latent space, and a decoder of similar complexity is used to map the latent representation to the final output, allowing for a complex relationship between the low-dimensional latent representation and the high-dimensional output. The simple linear decoders in ICNet, on the other hand, ensure that the latent representation in the bottleneck is constrained to directly reflect the dynamics that underlie neural activity. And because the bottleneck is shared across the individual decoder branches (Fig 3a), the bottleneck response $\hat{r}_b$ in ICNet reflects a general set of dynamics that is common to all animals and can be used on its own as a compact representation of central auditory processing. As a demonstration, we used the ICNet sound encoder as an acoustic front-end and trained a simple automatic-speech-recognition (ASR) DNN to classify phonemes from the bottleneck response $\hat{r}_b$. We found that the ASR system trained on the ICNet bottleneck response achieved similar performance to that of a baseline system trained with a Mel spectrogram input (Fig 3e). We also found that training on the ICNet bottleneck response resulted in better performance than training the same ASR system with the bottleneck responses from the nine single-branch models. This difference was small overall, but increased at relatively high signal-to-noise ratios (SNRs) (Fig 3f; average improvement of 1.24% [1.02, 1.76] for consonant recognition and 1.60% [1.25, 2.22] for vowel recognition across all SNRs and animals). These results demonstrate that the latent representation of ICNet can serve in the same manner as widely used acoustic front-ends such as auditory filterbanks or cochlear models for applications where simulation of central (rather than peripheral) auditory processing is desired.

ICNet is a highly accurate model of neural coding

Fig 4a shows example segments of recorded activity and predictions from ICNet for four sounds, with each image showing the response of 1000 units (the units were selected to achieve a uniform distribution of characteristic frequencies (CFs) on a logarithmic scale, reflecting the tonotopic organization of the IC). ICNet's simulations were nearly indistinguishable from the recordings across all units and all sounds. The full non-linearity of the mapping from sound to neural activity that is captured by ICNet is difficult to appreciate from visual inspection, but Fig 4a illustrates some of its remarkable features. For example, the speech in quiet elicited the strongest activity even though it was presented at the lowest intensity of any of the sounds (60 dB SPL with energy concentrated at low frequencies), while the speech in noise, which was presented at a much higher intensity (85 dB SPL with a much broader range of frequencies), elicited significantly less activity. The response to the moving ripples was particularly complex: the sound was presented at high intensity (85 dB SPL) with all energy concentrated at high frequencies, yet it elicited relatively weak activity in units with high CFs and relatively strong activity in units with low CFs. This frequency transposition indicates a high degree of non-linearity that was captured well by ICNet.

To provide an overall measure of the predictive power of ICNet, we used seven of the nine animals for which we had recordings of responses to the same sounds on successive trials. We computed the

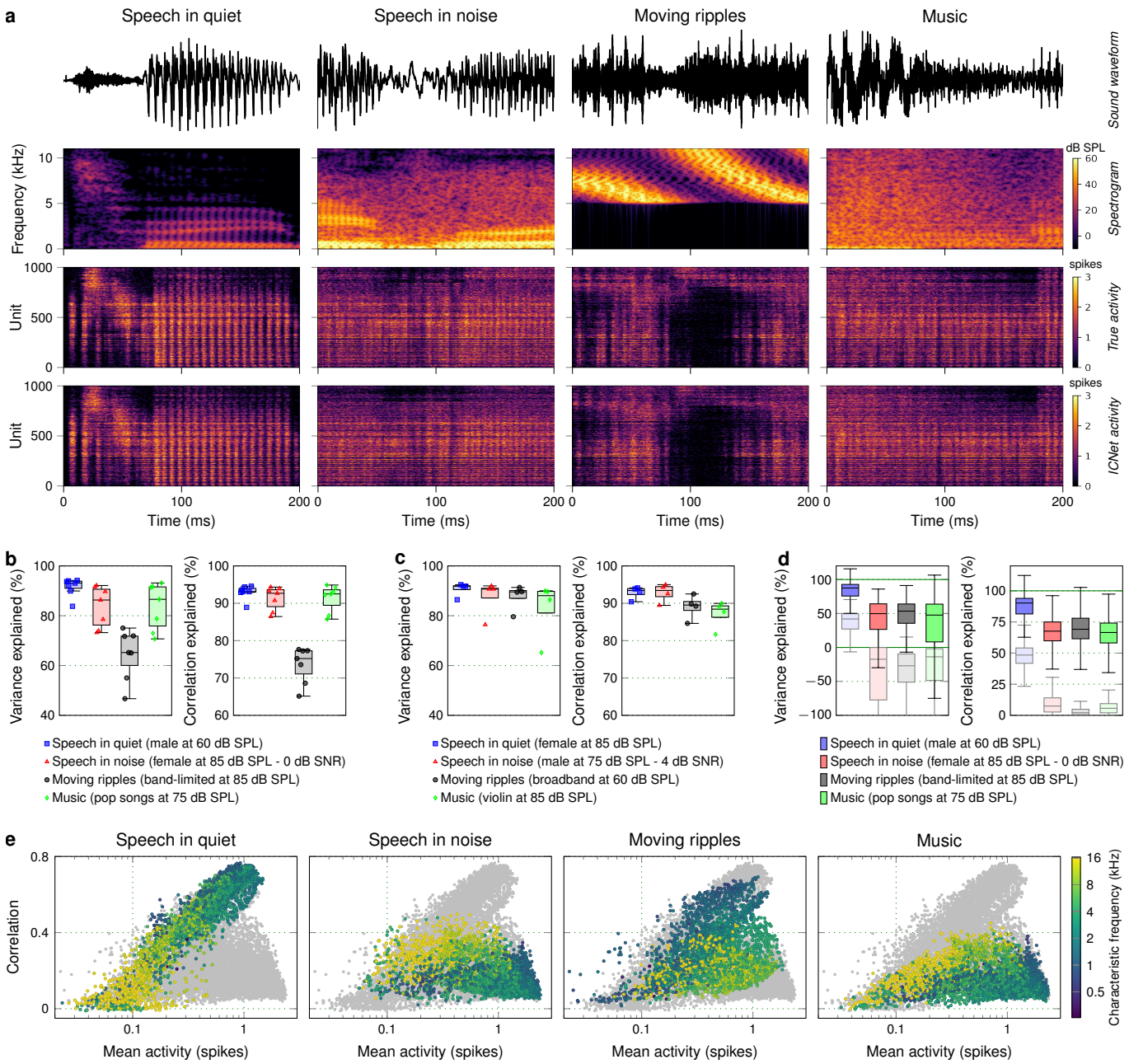

Fig 4. ICNet performance across sounds. a. Segments of recorded and ICNet predicted neural activity in response to 4 sounds. Each column shows the sound waveform and corresponding spectrogram, as well as the neural responses of 1000 units. Units were selected from all animals to achieve a uniform logarithmic spacing of CFs between 300 and 12000 Hz. Brighter colors indicate higher stimulus intensity for the spectrograms and higher activity for the neural responses. **b.** Predictive power of ICNet across 7 animals and 4 sounds, computed as the fraction of explainable variance and correlation explained by the model across all time bins and units of each animal. **c.** Predictive power of ICNet across 4 animals and 4 sounds, computed as the fraction of explainable variance and correlation explained by the model across all time bins and units of each animal. **d.** Predictive power of ICNet across 4 sounds and 3476 recorded units (from 7 animals), computed as the fraction of explainable variance and correlation explained by the model across all time bins of each unit. The shaded box plots show the predictive power of linear-nonlinear Poisson (LNP) models for the same units. **e.** Overall activity (mean spiking activity) is compared to reliability (correlation between two successive trials) of all 3476 units from 7 animals. Each panel shows the results for all units in response to one sound, with each point colored based on the CF of the unit. The gray points in the background show the results for all four sounds and are the same in each subplot.

fractions of the explainable variance and correlation (the covariance and correlation of recorded responses on successive trials) that ICNet explained for sounds that were not included during training (measured after flattening the responses by collapsing across time and units for each animal; see Methods). We found that ICNet performed best for speech in quiet, explaining more than 90% of the explainable variance and correlation across all animals (Fig 4b; 91.5% [91.4, 91.8] for variance and 92.9% [92.9, 93.1] for correlation). Performance was also excellent for speech in noise (83.6% [83.3 84.3] for variance, 91.2% [91.1 91.6] for correlation) and for music (83.6% [83.3 84.3] for variance, 91.3% [91.2 91.7] for correlation), and decreased only for the band-limited moving ripples (64.4% [64.0, 65.2] for variance, 73.5% [73.3, 74.0] for correlation). To further test the generality of ICNet, we also measured performance for four additional sounds that were presented on successive trials for four of the nine animals: speech in quiet at a higher intensity, speech in noise at a higher signal-to-noise ratio (SNR), broadband moving ripples, and a solo violin recording. We found that ICNet again explained most of the explainable variance and correlation for all sounds (Fig 4c; median values for variance and correlation between 88% and 93%).

The flattened metrics computed over all units are appropriate for assessing the overall predictive power of the model, but are dominated by the units with the most explainable variance or correlation (which are, presumably, the most important units to model accurately because their activity carries the most information about sound). When we instead examined ICNet's performance across individual units, we found lower values overall and substantial variability across units and sounds (Fig 4d). The lower values are a result of the difference in the way the flattened and unit-by-unit metrics account for the variation in mean activity across units for a given sound. This variation is fully accounted for by the flattened metrics but is either partly (for variance) or fully (for correlation) ignored when the metrics are computed unit by unit. This variation makes up a substantial portion of the explainable variance and correlation in the full neural activity and is well captured by the model, so the resulting performance values for the flattened metrics are higher than those for the individual units. ICNet outperformed standard single-layer linear-nonlinear Poisson (LNP) models of the same units by a wide margin (shaded box plots in Fig 4d), especially for non-speech sounds. (Note that negative values for the variance explained by LNP models indicate that the variance of the error in the model prediction was larger than the explainable variance in the response (Eq *4*).)

The variability in performance across units could have many sources. From the responses in Fig 4a, it is clear that units can vary in their activity depending on the match between the acoustic properties of a sound and the features to which a unit is responsive. Both overall activity (measured as mean count) and reliability (measured as the correlation of responses on successive trials) of individual units spanned a wide range that varied across sounds (Fig 4e), with a clear dependence on CF. To determine whether any of these properties (overall activity, reliability, preferred frequency) could explain the variation in model performance across units, we measured their correlation with the fraction of explainable correlation explained across all sounds. We found only a weak correlation between overall activity and model performance (-0.12 [-0.14, -0.10] for 3406 units from seven animals, after excluding units with the top and bottom 1% of explainable correlation explained) and no correlation between CF and model performance (-0.02 [-0.03, 0]). But the correlation between reliability and model performance was much stronger (0.54 [0.53, 0.55]). This suggests that the reliability of neural responses to a given sound is a strong determinant of ICNet performance.

Discussion

In this study, we combined large-scale neural recordings and deep learning to develop ICNet, a comprehensive model of neural coding in the early central auditory pathway. We presented a novel multi-branch encoder-decoder framework to describe the common features of sound encoding across a set of normal-hearing animals, while also accounting for the idiosyncrasy and non-stationarity of individual recordings. ICNet captures the statistical properties of neural activity with high resolution, simulating response patterns

across thousands of units with millisecond precision. The simulated responses capture most of the explainable variance in neural activity across a wide range of sounds. And because ICNet is based on convolutional operations, it is computationally fast: neural activity can be simulated more than 10x faster than real time, even on a CPU (for a 420 ms sound input, simulation of 512 units takes 49 ms on an Intel Xeon w7-2495X CPU and 39 ms on an NVidia RTX 4090 GPU).

In order to produce a model that would generalize well, we assembled a broad training dataset that contained representative samples of many complex sounds. We demonstrated that ICNet produced near-perfect simulations of the neural responses to a range of natural and artificial evaluation sounds that were not included in the training set (Fig 4a), capturing more than 90% of the explainable variance and correlation in recorded activity for speech in quiet (across all units and animals) and at least 86% of the explainable variance and correlation for speech in noise, music, and broadband moving ripples. The validity of ICNet as an accurate description of the auditory system could be further evaluated with respect to specific physiological phenomena that are traditionally used to characterize auditory processing for individual neurons (e.g. phase-locking [9], amplitude modulation tuning [4, 26–28] or spectro-temporal sensitivity [29]). This evaluation could be performed on the current model with a small additional dataset (to provide a reference for the manifestation of these phenomena in the IC) and the results could be used to inform the design of an updated training dataset for a new iteration of the model if needed.

Challenges in modeling neural coding

Accurate simulation of any system requires an appropriately specified model. The Poisson assumption that is typical for models of neural coding is often a poor match for the true distribution of activity, particularly in sub-cortical areas where activity can be strongly underdispersed [24, 30]. ICNet demonstrates that it is possible to model the full distribution of MUA counts without any statistical assumptions. We trained ICNet to infer non-parametric conditional probability distributions using a crossentropy loss which should, in principle, result in network weights that maximize the likelihood of the observed activity given the model. But since the prevalence of the different classes is imbalanced (58.4%, 23.2%, 13.7%, 4.2%, and 0.5% for MUA counts from 0 to 4, respectively, across our dataset), there is a risk that model performance will be biased against the under-represented classes. This risk can be addressed by training with a balanced loss, which weighs each class according to its prevalence in the dataset, or a focal loss, which weighs each class based on its predicted probability. The performance of ICNet did not improve significantly after training with a balanced or a focal crossentropy loss (results not shown), but explicit consideration of class imbalance may still prove beneficial when training models of neural coding for other brain areas.

In our datasets, non-stationarity was also a prominent feature due primarily to variations in brain state over the long duration of the recordings. Because our goal was to develop a generic model, we designed ICNet so that the effects of non-stationarity could be separated from the shared sound encoding and individualized for each animal. After training, ICNet can be used to provide a stationary simulation of neural activity across all animals by setting the individual time inputs to a fixed value. We recommend a value corresponding to a time near the beginning of recording, e.g., 1 hour, based on the assumption that the state of all animals is comparable at the beginning of recording and then diverges over time due to variations in the cumulative effects of the anesthesia. We chose to provide the model with a simple measure of time rather than a relevant biomarker (e.g., an EEG signature of anesthetic depth) because our reason for accounting for non-stationarity was ultimately to remove it. If the goal was instead to simulate the modulation of sensory coding by an external variable of interest, a measurement of this variable could be provided to the model in place of the time input. For example, to model the dependence of sensory processing on arousal, a measurement of pupil diameter could be used. Once trained, the resulting model could then, in principle, be used to simulate the response to any stimulus at any level of arousal and to probe the effects of arousal on the neural response.

We also designed ICNet with a shared encoder to capture the common features of sound encoding across the individual animals on which it was trained, as most applications of auditory models require a generic simulation of the fundamental features of auditory processing that are common across all normal-hearing individuals. Phenomenological or mechanistic models that are designed by hand will naturally have this property, but models that are learned directly from experimental data will inevitably reflect the idiosyncrasies of the individuals from which the data were collected. The results of our cross-validation experiments (Fig 3c) suggest that our ICNet model is generic, i.e., that the inclusion of additional animals during training would not improve its generality. With its limited dimensionality, ICNet's bottleneck layer thus provides a compact representation of sound encoding that can be used on its own for applications that require a general auditory model rather than explicit neural simulation. We demonstrated that this representation can be effective as a front-end for speech recognition (Fig 3e), and we expect that even better performance could be achieved in combination with DNN models with higher capacity. Depending on the needs of the application, ICNet's bottleneck could be modified to have specific properties, e.g., smoothness or orthogonality could be imposed by making the framework variational or by using more constrained decoder layers (for example, orthonormal or affine readouts [15]). More principled approaches to choosing hyperparameters such as Bayesian optimization might also be helpful to increase predictive power without significantly increasing model complexity.

Possible extensions and applications

We developed ICNet to describe MUA rather than the activity of single units, which is sufficient for many applications. But if simulations of single-unit activity are required, it should be possible to extend ICNet to produce them simply by retraining the decoder with a relatively small dataset of single-unit recordings. It should also be straightforward to extend ICNet to simulate awake activity, as we used anesthesia to facilitate the collection of the large-scale recordings required for initial model training. But with the trained ICNet as a starting point, it should be possible to learn the mapping from ICNet's latent representation to awake activity with a much smaller dataset. Another critical area for further development is binaural processing. We designed ICNet to capture the mapping between a single sensory input (a diotic stimulus in which the same sound was presented in both ears) and the neural activity recorded from both hemispheres of the IC. But IC activity is strongly modulated by binaural disparity, and these modulations encode information that is critical for spatial hearing. It is likely that a simple modification to ICNet that allows for two channels of sound input would allow it to capture the full complexity of binaural processing, but the new model would need to be retrained on a dichotic sound dataset.

We also designed ICNet so that it can be used either as a faithful simulator of neural coding in the IC or as a general model of auditory processing for sound engineering applications. Thus, it can simulate either response patterns from a large population of neural units or the latent neural dynamics that underlie these patterns. For hearing research, ICNet opens new possibilities with potential benefits that are both scientific (studies are no longer data limited) and ethical (animal experiments can be limited to confirmatory studies). ICNet has the potential to be a valuable tool for understanding the cortical mechanisms that facilitate stream segregation and auditory scene analysis, allowing downstream computations to be referenced directly to early neural representations. Cochlear models are already employed widely and successfully in this way [31, 32]; using ICNet as a second reference point midway along the central auditory pathway would enable early and late central processing to be decoupled.

Finally, ICNet provides a comprehensive normal-hearing reference for technology development. It can be used as a front-end in the design of systems that achieve natural (rather than objectively optimal) performance, as its outputs reflect the physiological constraints that are faced by the brain. The use of ICNet for optimization of signal processing (e.g., noise suppression) could unlock additional perceptual benefit beyond that achieved through optimization based directly on acoustics. The same may also be true for

the development of assistive listening technologies, with the output of ICNet serving as a target reference against which the efficacy of different hearing restoration strategies can be compared.

Methods

Experimental protocol

Experiments were performed on nine young-adult gerbils of both sexes that were born and raised in standard laboratory conditions. IC recordings were made at an age of 20-24 weeks. All experimental protocols were approved by the UK Home Office (PPL P56840C21).

Preparation for IC recordings

Recordings were made using the same procedures as in previous studies [20, 21]. Animals were placed in a sound-attenuated chamber and anesthetized for surgery with an initial injection of 1 ml per 100 g body weight of ketamine (100 mg per ml), xylazine (20 mg per ml), and saline in a ratio of 5:1:19. The same solution was infused continuously during recording at a rate of approximately 2.2 µl per min. Internal temperature was monitored and maintained at 38.7° C. A small metal rod was mounted on the skull and used to secure the head of the animal in a stereotaxic device. The pinnae were removed and speakers (Etymotic ER-10X) coupled to tubes were inserted into both ear canals. Sounds were high-pass filtered at 12 kHz (except for tones) and presented at 44.1 kHz without any filtering to compensate for speaker properties or ear canal acoustics. Two craniotomies were made along with incisions in the dura mater, and a 256-channel multi-electrode array was inserted into the central nucleus of the IC in each hemisphere.

Multi-unit activity

Neural activity was recorded at 20 kHz. MUA was measured from recordings on each channel of the electrode array as follows: (1) a bandpass filter was applied with cutoff frequencies of 700 and 5000 Hz; (2) the standard deviation of the background noise in the bandpass-filtered signal was estimated as the median absolute deviation / 0.6745 (this estimate is more robust to outlier values, e.g., neural spikes, than direct calculation); (3) times at which the bandpass filtered signal made a positive crossing of a threshold of 3.5 standard deviations were identified and grouped into 1.3 ms bins.

Characteristic frequency analysis

To determine the preferred frequency of the recorded units, 50 ms tones were presented with frequencies ranging from 294 Hz to 16384 Hz in 0.2 octave steps (without any high-pass filtering) and intensities ranging from 4 dB SPL to 85 dB SPL in 9 dB steps. Tones were presented 8 times each in random order with 10 ms cosine on and off ramps and 75 ms pause between tones. The characteristic frequency (CF) of each unit was defined as the frequency that elicited a significant response at the lowest intensity. Significance was defined as $p < 0.0001$ for the total MUA count elicited by a tone given an assumed Poisson distribution with rate equal to the higher of either the mean count observed in time windows of the same duration in the absence of sound or 1/8.

Deep neural network models

Time-invariant single-branch architecture

The time-invariant single-branch DNN that was used to predict the neural activity of individual recordings is shown in Fig 1. The model was trained to map a sound waveform s to neural activity $\hat{R}$ via an inferred conditional distribution of MUA counts $p(\hat{R}|s)$ using the following architecture: (1) a SincNet layer [33] with 48 bandpass filters of size 64 and stride of 1, followed by a symmetric logarithmic activation $y = \mathrm{sgn}(x) * \log(|x|+1)$; (2) a stack of five 1-D convolutional layers with 128 filters of size 64 and stride of 2, each followed by a PReLU activation; (3) a 1-D bottleneck convolutional layer with 64 filters of size 64 and a stride of 1, followed by a PReLU activation; and (4) a decoder convolutional layer without bias and $512 \times N_c$ filters of size 1, where N_c is the number of parameters in the count distribution $p(\hat{R}|s)$ as described below. All convolutional layers in the encoder included a bias term and used a causal kernel. A cropping layer was added after the bottleneck to remove the left context from the output, which eliminated any convolutional edge effects.

For the Poisson model, $N_c = 1$ and the decoder output was followed by a softplus activation to produce the rate function λ that defined the conditional count distribution for each unit in each time bin. (We also tested exponential activation but found that softplus yielded better performance.) The loss function was $\sum_{m=1}^{M}\sum_{t=1}^{T}(\lambda[m,t] - R[m,t]\log\lambda[m,t])$, where M and T are the number of units and time bins in the response, respectively. For the crossentropy model, $N_c = 5$ and the decoder output was followed by a softmax activation to produce the probability of each possible count $n \in \{0,...,N_c-1\}$ for each unit in each time bin. The loss function was $-\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{n=0}^{N_c-1} R[m,t]\log(p(\hat{R}[m,t]=n|s))$. The choice of 4 as the maximum possible count was made based on the MUA count distributions across all datasets (the percentage of counts above 4 was less than 0.02%). All neural activity was clipped during training and inference at a maximum value of 4.

Time-variant single-branch architecture

The time-variant single-branch architecture (Figs 2-3) included an additional time input τ. Each sample of the time input defined the relative time at which the corresponding sound sample was presented during a given recording (range of approximately 0-12 hr). To match the dimensionality and dynamic range of the sound input after the SincNet layer, the time input was given to a convolutional layer with 48 filters of size 1 (without bias), followed by a symmetric logarithmic activation. The resulting output was then concatenated to the output of the SincNet layer to form the input of the first convolutional layer of the encoder. The loss function was $-\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{n=0}^{N_c-1} R[m,t]\log(p(\hat{R}[m,t]=n|s,\tau))$.

Time-variant multi-branch architecture

The final ICNet model (Figs 3-4) comprises a time-variant DNN with a shared encoder and separate decoder branches for each of the individual animals. The architecture of the encoder and decoders were identical to those in the single-branch models, except for the additional time input. Instead of a single time input to the encoder, a separate time input was used for each animal with the same dimensionality and sampling rate as the bottleneck output. Each sample of the time input thus defined the relative time at which a corresponding MUA sample was recorded, and was given to a convolutional layer with 64 filters of size 1 (without bias), followed by a PReLU activation. The output of each of these layers was then multiplied (elementwise) by the bottleneck output and was given to the corresponding decoder.

Linear-nonlinear models

For the comparison of ICNet with linear-nonlinear models in Fig 3d, we trained single-layer models for 7 animals through Poisson regression. The models comprised a convolutional layer with 512 filters (one for each output unit) of size 64 samples and stride 1 (without bias), followed by an exponential activation. Mel spectrograms were used as inputs and were computed as follows: (1) the short-time Fourier transform (STFT) of the audio inputs was computed using a window of 512 samples and a step size of 32 samples, resulting in spectrograms with time resolution of 762.9395 Hz (to match the 1.3ms time bins of the neural activity); (2) the power spectrograms (squared magnitude of the STFT outputs) were mapped to 96 Mel bands with frequencies logarithmically spaced between 50 Hz to 12 kHz; (3) the logarithm of the Mel spectrograms was computed and was offset by 7 dB so that the spectrograms comprised only positive values; (4) a sigmoid activation function was applied to normalize the range of the spectrograms. An L2 regularization penalty was additionally applied to the convolutional kernel during training using a factor of 10^{-5}. We tested a number of parameters for the spectrogram inputs (STFT window size, linear or Mel scale, number of frequency bands, activation functions) and used the values that produced the best performance.

Training

Models were trained to transform 24414.0625 Hz sound input frames of 8192 samples into 762.9395 Hz neural activity frames of 256 samples (Fig 1b). Context of 2048 samples was added on the left side of the sound input frame (total of 10240 samples) and was cropped after the bottleneck layer (2048 divided by a decimation factor of 2^5 resulted in 64 cropped samples at the level of the bottleneck). Sound inputs were scaled such that an RMS of 0.04 corresponded to a level of 94 dB SPL. For the time-variant models, time inputs were expressed in seconds and were scaled by 1/36000 to match the dynamic range to that of the sound inputs.

The DNN models were trained on NVidia RTX 4090 GPUs using Python and Tensorflow [34]. A batch size of 50 was used with the Adam optimizer and a starting learning rate of 0.0004. All trainings were performed so that the learning rate was halved if the loss in the validation set did not decrease for 2 consecutive epochs. Early stopping was used and determined the total number of training epochs if the validation loss did not decrease for 4 consecutive epochs. This resulted in an average of 38.1 $\pm$ 3.4 epochs for the nine single-branch models, and a total of 42 epochs for the nine-branch ICNet model. To speed up training, the weights from one of the single-branch models were used to initialize the encoder of the nine-branch model.

Training dataset

All sounds that were used for training the DNN models totaled to 7.83 hr and are described below. 10% of the sounds were randomly chosen to form the validation set during training. The order of the sound presentation was randomized for each animal during recording.

Speech Sentences were taken from the TIMIT corpus [35] that contains speech read by a wide range of American English talkers. The entire corpus excluding “SA” sentences was used and was presented either in quiet or with background noise. The intensity for each sentence was chosen at random from 45, 55, 65, 75, 85, or 95 dB SPL. The speech-to-noise ratio (SNR) was chosen at random from either 0 or 10 when the speech intensity was 55 or 65 dB SPL (as is typical of a quiet setting such as a home or an office) or –10 or 0 when the speech intensity was 75 or 85 dB SPL (as is typical of a noisy setting such as a pub). The total duration of speech in quiet was 1.25 hr and the total duration of speech in noise was 1.58 hr.

Noise Background noise sounds were taken from the Microsoft Scalable Noisy Speech Dataset [36], which includes recordings of environmental sounds from a large number of different settings (e.g., café, office, roadside) and specific noises (e.g., washer-dryer, copy machine, public address announcements). The intensity of the noise presented with each sentence was determined by the intensity of the speech and the SNR as described above.

Processed speech Speech from the TIMIT corpus was also processed in several ways: (1) the speed was increased by a factor of 2 (via simple resampling without pitch correction of any other additional processing); (2) linear multi-channel amplification was applied, with channels centered at 0.5, 1, 2, 4, and 8 kHz and gains of 3, 10, 17, 22, and 25 dB SPL, respectively; or (3) the speed was increased and linear amplification was applied. The total duration of processed speech was 2.66 hr.

Music Pop music was taken from the musdb18 dataset [37], which contains music in full mixed form as well as in stem form with isolated tracks for drums, bass, vocals and other (e.g., guitar, keyboard). The total duration of the music presented from this dataset was 1.28 hr. Classical music was taken from the musopen dataset (`https://musopen.org`; including piano, violin and orchestral pieces) and was presented either in its original form; after its speed was increased by a factor of 2 or 3; after it was high-pass filtered with a cutoff frequency of 6 kHz; or after its speed was increased and it was high-pass filtered. The total duration of the music presented from this dataset was 0.45 hr.

Moving ripples Dynamic moving ripple sounds were created by modulating a series of sustained sinusoids to achieve a desired distribution of instantaneous amplitude and frequency modulations. The lowest frequency sinusoid was either 300 Hz, 4.7 kHz, or 6.2 kHz. The highest frequency sinusoid was always 10.8 kHz. The series contained sinusoids at frequencies between the lowest and the highest in steps of 0.02 octaves, with the phase of each sinusoid chosen randomly from between 0 and 2π. The modulation envelope was designed so that the instantaneous frequency modulations ranged from 0 to 4 cycles/octave, the instantaneous amplitude modulations ranged from 0 to 10 Hz, and the modulation depth was 50 dB. The total duration of the ripples was 0.22 hr.

Evaluation

To predict neural activity using the trained DNN models, the decoder parameters were used to define either a Poisson distribution or a categorical probability distribution with N_c classes using the Tensorflow Probability toolbox [38]. The distribution was then sampled from to yield simulated neural activity across time bins and units, as shown in Fig 1b.

Evaluation dataset

All model evaluations (Figs 1-4) used only sounds that were not part of the training dataset. Each sound segment was 30 s in duration. The timing of the presentation of the evaluation sounds varied across animals. For two animals, each sound was presented only twice and the presentation times were random. For three animals, each sound was presented four times, twice in successive trials during the first half of the recording and twice in successive trials during the second half of the recording. For the other four animals, each sound was presented ten times, twice in successive trials at times that were approximately 20, 35, 50, 65, and 80% of the total recording time, respectively.

Speech in quiet For all animals, a speech segment from the UCL SCRIBE dataset (`http://www.phon.ucl.ac.uk/resource/scribe`) consisting of sentences spoken by a male talker was presented at 60 dB SPL. For four of the animals (Fig 4b), another speech segment from the same dataset was chosen and was presented at 85 dB SPL, consisting of sentences spoken by a female talker.

Speech in noise For all animals, a speech segment from the UCL SCRIBE dataset consisting of sentences spoken by a female talker was presented at 85 dB SPL in hallway noise from the Microsoft Scalable Noisy Speech dataset at 0 dB SNR. For four of the animals (Fig 4b), another speech segment from the same dataset was chosen and was presented at 75 dB SPL, consisting of sentences spoken by a male talker mixed with café noise from the Microsoft Scalable Noisy Speech dataset at 4 dB SNR.

Moving ripples For all animals, dynamic moving ripples with the lowest frequency sinusoid of 4.7 kHz were presented at 85 dB SPL. For four of the animals (Fig 4b), dynamic moving ripples with the lowest frequency sinusoid of 300 Hz were also used and were presented at 60 dB SPL.

Music For all animals, three seconds from each of 10 mixed pop songs from the musdb18 dataset were presented at 75 dB SPL. For four of the animals (Fig 4b), a solo violin recording from the musopen dataset was presented at 85 dB SPL.

Pure tones For measuring Fano factors (Fig 1), 75 ms tones at 4 frequencies (891.44, 2048, 3565.78 and 8192 Hz) were presented at 85 dB SPL with 10 ms cosine on and off ramps. Each tone was repeated 128 times, and the resulting responses were used to compute the average, the variance and the full distribution of MUA counts (using time bins from 2.5 ms to 80 ms).

Non-stationarity

For Fig 2g, we quantified the non-stationarity in each recording as the absolute difference of the covariance of neural activity elicited by the same sound on two successive trials and on two trials recorded several hours apart. The covariance was computed after flattening (collapsing neural responses across time and units into one dimension). To obtain a single value for the covariance on successive trials, the values from the two pairs of successive trials (one pair from early in the recording and one pair from late in the recording) were averaged. To obtain a single value for the covariance on separated trials, the values from the two pairs of separated trials were averaged.

Overall performance metrics

RMSE and log-likelihood were used as overall performance metrics (Figs 1-3). The reported results for both metrics were the averages computed over two trials of each sound within each individual recording. One trial was chosen from the first half of the recording and the other from the second, resulting in times of 3.51 $\pm$ 1.83 hr and 8.92 $\pm$ 1.16 hr across all nine animals. For each trial, the RMSE and log-likelihood were computed as follows:

$$\text{RMSE} = \sqrt{\frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}(R[m,t]-\hat{R}[m,t])^2} \tag{1}$$

$$\text{Log-likelihood} = \frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}\log(p(\hat{R}[m,t]=R[m,t]|s,\tau)), \tag{2}$$

where M is the number of units, T is the number of time bins, s is the sound stimulus, τ is the time input, R is the recorded neural activity, and $\hat{R}$ is the predicted neural activity obtained by sampling from the inferred distribution $p(\hat{R}|s,\tau)$ (or $p(\hat{R}|s)$ for a time-invariant model).

Predictive power metrics

To assess the predictive power of ICNet in Fig 4, we selected two metrics that were computed using successive trials for the seven animals for which this was possible (chosen from the beginning of the recording, resulting in times of 1.94 $\pm$ 0.46 hr across animals). The two metrics are formulated as the fraction of explainable variance or correlation that the model explained across all units. The fraction of explainable correlation explained was computed as follows:

$$\text{Correlation explained (\%)} = \frac{100}{2} \cdot \frac{\rho(R_1,\hat{R}_1)+\rho(R_2,\hat{R}_2)}{\rho(R_1,R_2)}, \tag{3}$$

where $R_1, R_2 \in \mathbb{N}^{M\times T}$ are the recorded neural activity for each trial, $\hat{R}_1, \hat{R}_2 \in \mathbb{N}^{M\times T}$ are the predicted neural activity, and ρ is the correlation coefficient. The responses were flattened (collapsed across time and units into one dimension) before the correlation was computed.

The fraction of explainable variance explained was computed as follows [14]:

$$\text{Variance explained (\%)} = 100 \cdot \left(1 - \frac{\frac{1}{2MT}\sum_{m=1}^{M}\sum_{t=1}^{T}(R_1 - E_c[\hat{R}_1])^2 + \frac{1}{2MT}\sum_{m=1}^{M}\sum_{t=1}^{T}(R_2 - E_c[\hat{R}_2])^2 - \sigma^2_{noise}}{\frac{1}{2}(Var[R_1]+Var[R_2]) - \sigma^2_{noise}}\right), \tag{4}$$

$$\sigma^2_{noise} = \frac{1}{2}Var[R_1 - R_2], \tag{5}$$

where E_c denotes expectation over counts of $p(\hat{R}|s,\tau)$ (or $p(\hat{R}|s)$ for a time-invariant model). The responses were flattened (collapsed across time and units into one dimension) before the variances in the denominator of Eq *4* and in Eq *5* were computed, while flattening was achieved in the numerator of Eq *4* through the averaging across units and time. To compute the predictive power of ICNet for each unit (Fig 4d), we computed the formulas of Eqs *3*-*5* across time only (without collapsing across the unit dimension or averaging across units). Values above 100% for individual units indicate that the ICNet responses were more similar to the recorded responses than the recorded responses were to each other across successive trials, which is possible in recordings with high non-stationarity.

Phoneme recognition

A DNN architecture based on Conv-TasNet [39] was used to train an automatic-speech-recognition (ASR) back-end that predicted phonemes from the ICNet bottleneck response. The DNN architecture comprised: (1) a convolutional layer with 64 filters of size 3 and no bias, followed by a PReLU activation, (2) a normalization layer, followed by a convolutional layer with 128 filters of size 1, (3) a block of 8 dilated convolutional layers (dilation from 1 to 2^7) with 128 filters of size 3, including PReLU activations and residual skip connections in between, (4) a convolutional layer with 256 filters of size 1, followed by a sigmoid activation, (5) an output convolutional layer with 40 filters of size 3, followed by a softmax activation. All convolutional layers were 1-D and used a causal kernel with a stride of 1.

The ASR model was trained using the train subset of the TIMIT corpus [35]. The TIMIT sentences were resampled to 24414.0625 Hz and were segmented into sound input frames of 81920 samples (windows of 65536 samples with 50% overlap and left context of 16384 samples). In each training step, the sound input frames were given to the frozen ICNet model to generate bottleneck responses at 762.9395 Hz (output frames of 2048 samples with left context of 256 samples). The bottleneck response was then given to the

ASR back-end to predict the probabilities of the 40 phoneme classes across time (output frames of 2048 samples at 762.9395 Hz after cropping the context).

The TIMIT sentences were calibrated to randomly selected levels between 40 and 90 dB SPL in steps of 5 dB, and were randomly mixed with the 18 noise types from the DEMAND dataset [40] at SNRs of -30, -20, -10, 0, 10, 20, 30, and 100 dB. The phonetic transcriptions of the TIMIT dataset were downsampled to the sampling frequency of the neural activity (762.9395 Hz) and were grouped into 40 classes (15 vowels and 24 consonants plus the glottal stop). To account for the phonetic class imbalance (high prevalence of silence in the dataset), a focal crossentropy loss function was used for training with the focusing parameter γ set to 5.

ASR models were trained to predict speech recognition from the ICNet bottleneck response, as well as from the bottleneck responses of the nine single-branch time-variant models. The trained ASR models were evaluated using 6 random sentences of the test subset of the TIMIT corpus. The sentences were calibrated at 65 dB SPL, and speech-shaped noise was generated to match the long-term average spectrum of the sentences. A baseline system was also trained with the same DNN architecture, but using a standard Mel spectrogram as the input. The Mel spectrogram was computed from the sound input (24414.0625 Hz) using frames of 512 samples, a hop size of 32 samples and 64 Mel channels with frequencies logarithmically spaced from 50 to 12000 Hz. These parameters were chosen to maximize performance while matching the resolution of the ICNet bottleneck response (762.9395 Hz output frames with 64 channels).

Statistics

Confidence intervals for all reported values were computed using 1000 bootstrap samples. For the metrics that we used to assess model performance (RMSE, log-likelihood, variance and correlation explained), bootstrapping was performed across units for each animal. For the ASR results (Fig 3f), bootstrapping was performed across animals.

Data availability

All of the sounds and IC recordings that were used to evaluate ICNet in Fig 4b and d are available via `https://doi.org/10.5281/zenodo.12636772`. The four sounds can be used as inputs to ICNet to simulate responses for comparison with the IC recordings provided for seven animals. Researchers seeking access to the full set of neural recordings for research purposes should contact the corresponding author via e-mail to set up a material transfer agreement.

Code availability

The code of the ICNet model is available via `https://doi.org/10.5281/zenodo.11943090` or `https://github.com/fotisdr/ICNet`. A Jupyter notebook is included with a simple usage example for ICNet. A non-commercial, academic UCL license applies.

Acknowledgments

The authors thank Jennifer Bizley and Lloyd Pellatt for their advice.

Author contributions

F.D.: Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing: Original Draft Preparation; **S.S.**: Investigation; **Y.X.**: Investigation; **A.F.**: Conceptualization, Project administration, Software, Supervision; **N.A.L.**: Conceptualization, Data Curation, Funding acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing - Review & Editing

Competing interests

N. A. L. is a co-founder of Perceptual Technologies.

References

1. Osses Vecchi A, Varnet L, Carney LH, Dau T, Bruce IC, Verhulst S, et al. A comparative study of eight human auditory models of monaural processing. Acta Acustica. 2022;6:17.
2. Meyer AF, Williamson RS, Linden JF, Sahani M. Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation. Frontiers in Systems Neuroscience. 2017;10:109.
3. Rahman M, Willmore BD, King AJ, Harper NS. Simple transformations capture auditory input to cortex. Proceedings of the National Academy of Sciences. 2020;117(45):28442–28451.
4. Nelson PC, Carney LH. A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. The Journal of the Acoustical Society of America. 2004;116(4):2173–2186.
5. Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America. 2005;118(2):887–906.
6. Carney LH, Li T, McDonough JM. Speech coding in the brain: representation of vowel formants by midbrain neurons tuned to sound fluctuations. eNeuro. 2015;2(4).
7. Farhadi A, Jennings SG, Strickland EA, Carney LH. Subcortical auditory model including efferent dynamic gain control with inputs from cochlear nucleus and inferior colliculus. The Journal of the Acoustical Society of America. 2023;154(6):3644–3659.
8. Baby D, Van Den Broucke A, Verhulst S. A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. Nature Machine Intelligence. 2021;3(2):134–143.
9. Drakopoulos F, Baby D, Verhulst S. A convolutional neural-network framework for modelling auditory sensory cells and synapses. Communications Biology. 2021;4(1):827.
10. Nagathil A, Bruce IC. WaveNet-based approximation of a cochlear filtering and hair cell transduction model. The Journal of the Acoustical Society of America. 2023;154(1):191–202.
11. Van Den Broucke A, Drakopoulos F, Baby D, Verhulst S. Otoacoustic emissions in a deep-neural-network model of cochlear mechanics. In: The 14th International Mechanics of Hearing Workshop (MoH 2022). vol. 3062; 2024.

12. McIntosh L, Maheswaranathan N, Nayebi A, Ganguli S, Baccus S. Deep learning models of the retinal response to natural scenes. Advances in Neural Information Processing Systems. 2016;29.

13. Walker EY, Sinz FH, Cobos E, Muhammad T, Froudarakis E, Fahey PG, et al. Inception loops discover what excites neurons most using deep predictive models. Nature Neuroscience. 2019;22(12):2060–2065.

14. Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Computational Biology. 2019;15(4):e1006897.

15. Lurz KK, Bashiri M, Willeke K, Jagadish AK, Wang E, Walker EY, et al. Generalization in data-driven models of primary visual cortex. bioRxiv. 2020; p. 2020.10.05.326256.

16. Bashivan P, Kar K, DiCarlo JJ. Neural population control via deep image synthesis. Science. 2019;364(6439):eaav9436.

17. Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron. 2018;98(3):630–644.

18. Keshishian M, Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N. Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. eLife. 2020;9:e53445.

19. Pennington JR, David SV. A convolutional neural network provides a generalizable model of natural sound coding by neural populations in auditory cortex. PLoS Computational Biology. 2023;19(5):e1011110.

20. Sabesan S, Fragner A, Bench C, Drakopoulos F, Lesica NA. Large-scale electrophysiology and deep learning reveal distorted neural signal dynamics after hearing loss. eLife. 2023;12:e85108.

21. Armstrong AG, Lam CC, Sabesan S, Lesica NA. Compression and amplification algorithms in hearing aids impair the selectivity of neural responses to speech. Nature Biomedical Engineering. 2022;6(6):717–730.

22. Garcia-Lazaro JA, Belliveau LA, Lesica NA. Independent population coding of speech with sub-millisecond precision. Journal of Neuroscience. 2013;33(49):19362–19372.

23. Lyamzin DR, Garcia-Lazaro JA, Lesica NA. Analysis and modelling of variability and covariability of population spike trains across multiple time scales. Network: Computation in Neural Systems. 2012;23(1-2):76–103.

24. Pecka M, Siveke I, Grothe B, Lesica NA. Enhancement of ITD coding within the initial stages of the auditory pathway. Journal of Neurophysiology. 2010;103(1):38–46.

25. Williamson RS, Sahani M, Pillow JW. The equivalence of information-theoretic and likelihood-based methods for neural dimensionality reduction. PLoS Computational Biology. 2015;11(4):e1004141.

26. Koumura T, Terashima H, Furukawa S. Cascaded tuning to amplitude modulation for natural sound recognition. Journal of Neuroscience. 2019;39(28):5517–5533.

27. Kim DO, Carney L, Kuwada S. Amplitude modulation transfer functions reveal opposing populations within both the inferior colliculus and medial geniculate body. Journal of Neurophysiology. 2020;124(4):1198–1215.

28. Koumura T, Terashima H, Furukawa S. Human-like modulation sensitivity emerging through optimization to natural sound recognition. Journal of Neuroscience. 2023;43(21):3876–3894.

29. Khatami F, Escabí MA. Spiking network optimized for word recognition in noise predicts auditory system hierarchy. PLOS Computational Biology. 2020;16(6):e1007558.

30. Kara P, Reinagel P, Reid RC. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. Neuron. 2000;27(3):635–646.

31. Shan T, Cappelloni MS, Maddox RK. Subcortical responses to music and speech are alike while cortical responses diverge. Scientific Reports. 2024;14(1):789.

32. Saddler MR, McDermott JH. Models optimized for real-world tasks reveal the necessity of precise temporal coding in hearing. bioRxiv. 2024; p. 2024.04.21.590435.

33. Ravanelli M, Bengio Y. Speaker recognition from raw waveform with SincNet. In: 2018 IEEE Spoken Language Technology Workshop (SLT); 2018. p. 1021–1028.

34. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16); 2016. p. 265–283.

35. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon technical report n. 1993;93:27403.

36. Reddy CK, Beyrami E, Pool J, Cutler R, Srinivasan S, Gehrke J. A scalable noisy speech dataset and online subjective test framework. arXiv. 2019; p. 1909.08050.

37. Rafii Z, Liutkus A, Stöter FR, Mimilakis SI, Bittner R. The MUSDB18 corpus for music separation; 2017.

38. Dillon JV, Langmore I, Tran D, Brevdo E, Vasudevan S, Moore D, et al. Tensorflow distributions. arXiv. 2017; p. 1711.10604.

39. Luo Y, Mesgarani N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019;27(8):1256–1266.

40. Thiemann J, Ito N, Vincent E. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. In: Proceedings of Meetings on Acoustics. vol. 19; 2013.