Title: **Inferring the Neural Basis of Binaural Detection Using Deep Learning**

Authors: Samuel S. Smith*[1], Joseph Sollini[1], Michael A. Akeroyd[1]

1- Hearing Sciences, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, Nottingham, NG7 2RD, UK

*Correspondence: samuel.smith@nottingham.ac.uk

Keywords: binaural hearing, signal detection, deep learning, cross-correlation, psychophysics, neurophysiology

Word Count: Summary, 169; Results and Discussion, 2239; Methods, 1307

SUMMARY

Parallels have been reported between broad organization in the auditory system and optimized artificial neural networks[1–3]. It remains to be seen whether such promising analogies between the auditory system and deep learning models endure at other levels of description. Here, we examined whether artificial neural networks[4,5] could offer a mechanistic account of human behavior in an auditory task. The chosen task promoted the use of binaural cues (across the ears) to help detect a signal in noise[6,7]. In the optimal network, we observed the emergence of specialized computations with prominent similarities to *in vivo* animal data[8]. Artificial neurons developed a sensitivity to temporal delays that increased hierarchically, and were widely distributed in preference (extending to delays beyond the range permitted by head width). Ensuing dynamics were consistent with a binaural cross-correlation mechanism[9]. Given the neural mechanisms of binaural detection in humans is contested[9–13], these findings help to resolve this debate. Moreover, this is a primary demonstration that deep learning can infer tangible mechanisms underlying auditory perception.

RESULTS

Here, we propose that deep neural networks (DNNs) can be leveraged to learn about the neural mechanisms of hearing. We employed an autoencoder-based architecture that puts interpretability at its forefront (as have recently shown potential in the field of physics[4,5]). DNNs were trained to account for human behaviour in an auditory task, specifically, binaural sound detection[6,7] (a phenomenon where interaural differences enhance the detectability of one sound amongst another by up to 15 dB). The task was selected as it typifies a situation in which specialized neural mechanisms have developed to extract and utilize auditory information. We were interested in the capacity for DNNs to do likewise.

To augment the availability of training data, we generated data from a set of equations recognized as effective in predicting human binaural psychophysics[12,14–18]. DNNs were trained to mimic detection of a 500 Hz pure tone amongst a broadband noise, each with interaural time differences (ITDs) that varied trial-to-trial. The range of ITDs was restricted to fall within the human physiological range, i.e. as though they came from randomly chosen azimuthal locations in the real-world (**Figure 1A**). We found that the optimal DNN predicted unseen binaural detection data with a root mean square error of 2.5% (**Figure S1A**). The predictions and internal dynamics of this DNN are the focus of this report (summary analytics across other DNN configurations are shown in **Figure S1**).

Deep neural network accounts for human binaural detection psychophysics.

As expected, we found that the DNN's detection thresholds (i.e. 69% correct performance) decreased as ITD difference between tone and noise increased (**Figure 1C,D**). For example, in diotic noise (noise ITD = 0) where the tone came from the far left, detection thresholds were significantly enhanced by 9 dB (from a maximum of 30.8 dB to 21.8 dB, two-sided unpaired t-test, $p \ll 0.0001$). To allow comparative assessment of the DNN and previously published data, we also tested the network on stimulus configurations typically employed to study binaural detection. These include tones and noise configurations which are either in-phase or completely out-of-phase across the ears. In the literature these stimuli are denoted as NoSo, NoSπ, NπSπ, NπSo, where N refers to the noise component, S the pure tone signal, with the successive subscripts denoting interaural phase difference (IPD) in radians. These stimuli use ITDs that fall outside of the physiological range. For example, a 500 Hz pure tone with an IPD of π corresponds to an ITD of 1000 μs, larger than that produced by the head size in the simulated training data (maximum of 655 μs, calculated with Woodworth's formula[19]). This meant the DNN had no prior exposure to this size of ITD and so it was unclear how it would function over this range. We found that when the noise signal had zero IPD, the mean detection thresholds predicted by the DNN for corresponding homophasic (NoSo) and antiphasic (NoSπ) tones were 30.9 dB and 20.8 dB respectively. The gain in detection, commonly called the binaural masking level difference (BMLD), was 10.1 dB ($p \ll 0.0001$, two-sided unpaired t-test). Comparatively, when instead the noise signal was interaurally out of phase, the mean detection thresholds predicted for the corresponding homophasic (NπSπ) and antiphasic (NπSo) stimuli were 26.6 dB and 17.1 dB respectively. Their BMLD was 9.5 dB ($p \ll 0.0001$, two-sided unpaired t-test). These BMLDs are similar to those measured in people[20] and with estimates from psychophysical equations (10.7 dB and 10.3 dB respectively; **Figure 1E**).

Time delay tuning emerges early in the network.

The DNN was able to account for key aspects of human binaural detection behaviour. Given this, we next sought to understand the means by which the DNN derived this behaviour, i.e. the mechanism(s). A common property of animal binaural systems is ITD tuning (**Figure 2A**). We found that ITD tuning emerged hierarchically within the early layers of the network. To demonstrate this we characterized "noise delay" functions in DNN nodes, i.e. their response to noises presented with varying ITDs as typically measured in animal physiology studies[21]. For nodes in the DNN's first layer, we observed significant ITD tuning in 63 out of 100 nodes (**Figure 2B**). Tuning was quantified by regressing a Gabor function[22], the combination of a cosine windowed by a Gaussian (overlaid in **Figure 2D**). By the DNN's second layer, significant ITD tuning had emerged in all 100 nodes (**Figure 2E**). Estimates of each nodes' best ITD (bITD) were inferred from Gabor fits (in order to account for nodes that

were oscillatory in their noise delay responses, a form of phase-ambiguity noted in animal physiology[23], bITD was attributed to the most central tuning peak). In both the first and second layers of the DNN, we observed a wide distribution of bITDs, both within the simulated head-range, and beyond it.

Network dynamics match those of a cross-correlation mechanism.

ITD tuning is an axiomatic property in many explanations of binaural detection. Therefore, in previous animal studies[8,22,24,25], clarity regarding the mechanism underlying binaural detection has been obtained by examining neural responses to a set of commonly employed tone-in-noise conditions (i.e. NoSo, NoSπ, NπSπ, NπSo). To examine the mechanism inferred by the DNN, we measured node activations in responses to the same stimuli. We found that in the DNN's second layer, node responses varied by bITD and the interaural phase of the noise presented (two-way ANOVA, $F[98,202]=12.5$ for main effect of bITD, $F[1,202]=31.2$ for main effect of noise phase, $F[98,202]=2.9$ for their interaction, $p \ll 0.0001$ for all). When a signal was presented amongst an in-phase noise (No), responses were largest for nodes with bITDs near 0 μs and decreased as bITDs were increasingly non-zero. Conversely, amongst an out-of-phase noise (Nπ), responses were lowest for nodes with bITDs near 0 μs and increased as bITDs deviated away from this (**Figure 3A**). The effects of tone phase on node dynamics were more subtle, although these dynamics were also in accordance with a nodes' tuning properties. Nodes tuned to smaller ITDs responded most to in-phase tones (So) and least to out-of-phase tones (Sπ), and vice-versa for nodes tuned to larger ITDs (**Figure 3B,C**).

These responses are commensurate with a binaural cross-correlation mechanism. The concept of binaural cross-correlation is predicated on the existence of coincidence detectors that encode temporally offset signals (similar to the dynamics already described in layers 1 and 2)[26]. Computationally, a binaural cross-correlation can be calculated by summing the point-by-point product of two temporally offset signals. Comparative outputs from a simple binaural cross-correlation algorithm (namely for signals in noise passed through narrow-band filters centered at 500 Hz) are shown in **Figure 3D-F**, and were found to resemble responses across the DNN's layer 2 nodes (Pearson's $r=0.36$, $p \ll 0.0001$). A number of animal studies have reported neural responses consistent with a cross-correlation mechanism[8,22,24,25] (**Figure 3G-I**).

Early network ablation is detrimental to binaural detection.

The functional importance of early layers in the DNN was further interrogated by inflicting targeted damage and observing knock-on effects to BMLDs. We set to zero the weights of a fixed proportion of nodes in a specific layer of the network, i.e. ablating them. Our ablation range varied from 0% (none) to 50% (lots). We then observed the corresponding NoSo-NoSπ BMLDs (**Figure S2**). We found that ablation to nodes in the DNN's first layer were

most detrimental to BMLDs, to such an extent that with as little as 5% ablation the DNN failed to predict a significant BMLD (criterion set at $p<0.05$ with Bonferonni correction, Student's two-tailed t-test). When ablating the second layer, the DNN initially failed to predict a BMLD when 35% of nodes were ablated. Layers later in the network were more robust to the effects of ablation. BMLDs withstood ablations up to and including 50% (maximum tested) of the nodes in fourth layer, and significant BMLDs were not exhibited when 40% or more of the nodes in the fifth layer were ablated.

Low-dimensional network representations imitate neural signature of population-level cortical activity.

Responses in the DNN's second layer combine together to form the low-dimensional representations encoded in the DNN's central nodes (i.e. layer 3). This bottleneck architecture has proven successful in extracting key conceptual variables in other fields[4], and our ambition was for something analogous for binaural detection. In the DNN, six central nodes were deemed operational (**Figure S1B**), and we found noteworthy similarities between these nodes and population level cortical responses in the guinea pig[8] (comparable observations have also been reported in other animals[22,25] and in the guinea pig inferior colliculus[24]). The central nodes could take the value of any real number, positive or negative (a necessary limitation imposed by the network autoencoder architecture). Although not essential for our main conclusions, node responses are presented polarity-corrected to best correspond with noise ITD tuning in cortex (**Figure 4A,C**; guinea pig auditory cortex).

An interesting feature of binaural processing can be seen when comparing detection behaviour and animal physiological data. Similar improvements in behavioural performance can be attributed to completely different alterations in neural dynamics. For example, BMLDs across NoSo-NoSπ conditions are similar to those found across NπSπ-NπSo conditions (**Figure 1E**). However, guinea pig neural data[8] suggests different neural dynamics underlie these similar BMLDs. In cortical recordings, population spike counts dropped amongst an No signal, as a pure tone went from So to Sπ (left panel in **Figure 4B**). Conversely, amongst an Nπ signal, as a pure tone transitioned from Sπ to So, population spike counts increased (right panel in **Figure 4B**). These opposing dynamics therefore represent a unique signature of processing

The majority of layer 3 nodes displayed the same opposing dynamics in response to homophasic/antiphasic stimuli as observed in the guinea pig auditory cortex. Threshold responses in four latent nodes (n_1, n_3, n_4 and n_5, in left panels of **Figure 4D**) were found to be significantly lower in response to NoSπ relative to NoSo ($p<0.05$, two-sided unpaired t-test). Conversely, threshold responses in the same nodes to NπSo at a threshold level were significantly higher in comparison to NπSπ ($p<0.05$, two-sided unpaired t-test, right panels in **Figure 4D**). One latent node was qualitatively different to the others (node 6 of **Figure 4D**),

and seemed to encode offsets related to the interaural phase of noise (No v. Npi, $p \ll 0.0001$, two-sided unpaired t-test).

DISCUSSION

Deep neural networks as analogues of the auditory system.

Recent DNN studies have described similarities with human non-invasive imaging data[2] and with broad organizational principles in the auditory system (e.g. changes in temporal encoding)[1,3]. Here, we set out to discover the efficacy of DNNs as a means to infer more specific, system-level, auditory mechanisms. We modelled binaural hearing (namely, in the context of sound detection), representative of a highly specialized system for which the inner workings of DNN analogues have scarcely been considered[27–29]. Perhaps easy to overlook, our DNN was able to successfully utilize binaural discrepancies in auditory stimuli, as opposed to seeking an alternative strategy[30] and/or failing to exhibit binaural detection behaviour. Notably, similarities with animal single-unit and population-level neural responses were observed. ITD tuning, a well-known characteristic of binaural neurons, more often discussed in the context of sound localization[21] rather than detection, emerged early in the DNN. Additionally, the dynamics of second layer nodes, in response to tones presented in broadband noise, resembled a binaural cross-correlator[9]. Latent nodes, central within the DNN's architecture, were compatible with the downstream dynamics of a cross-correlation mechanism and resembled guinea pig cortical neural recordings[8]. These dynamics were not hard-coded into the network and together highlight insights obtainable by treating autoencoder-based DNNs as model organisms[31].

However, comparison between DNNs and neural biology come accompanied by an asterisk. The network was not constructed with the goal of accurately mimicking neuronal biophysics or hierarchical complexity, but instead a trade-off was made to favor mechanistic interpretation and optimization performance. The inclusion of additional structural priors (e.g. hemispherity) and biologically inspired processes (e.g. spiking neural networks[1,32]) could have merit, but any impact on interpretability should be carefully weighed.

The open problem of binaural detection in humans.

In addition to binaural detection serving as an ideal topic to interrogate DNN capabilities, it is a key function for successful communication in real-life environments. Consequently, binaural deficits have real-world consequences[33,34]. However, to date, understanding of binaural detection in humans has been mired due to ambiguity regarding whether animal neurophysiology data satisfactorily accounts for human psychophysics. There is ongoing debate as to which of a number of theoretical frameworks best relate to human binaural detection[9,10,12,13]. For example, animal neural data lend support to a theory of binaural

cross-correlation[8,22,24,25]. Whereas, human behaviour appears to be equally well, if not better, described by a noise equalization and subtraction scheme[12,16,17]. These discrepancies have not been resolved with human imaging data[35–38], for which resolution and response variability are key limitations[39]. Our results indicate that, in a highly connected network, binaural cross-correlation functionality is favoured to explain human psychophysics. This preference is only reinforced by the fact that the equations used to generate training data were derived from a theory of noise equalization and subtraction[12,18] (although it should be noted that, when described in purely statistical terms, both strategies are indiscriminate at threshold[10]). DNNs offer a means to bridge the gap between animal and human data.

Conclusion.

In conclusion, our results indicate that an artificial neural network can discover specialized binaural mechanisms in order to explain human binaural detection. These mechanisms (namely, temporal delay tuning followed by a cross-correlator) corroborate observations made in animal physiology. The work demonstrates the potential for deep learning, in unison with experimental data, to clarify human auditory perception.

ACKNOWLEDGMENTS

S.S.S. and M.A.A. are supported by the Medical Research Council (grant number MR/S002898/1). J.S. is funded by a Nottingham Research Fellowship from the University of Nottingham. We are grateful for access to the University of Nottingham's Augusta HPC service. Thanks to Alan Palmer for assistance in sharing previously published physiology data.

AUTHOR CONTRIBUTIONS

Conceptualization, S.S.S. and M.A.A.; Methodology, S.S.S.; Investigation, S.S.S. and J.S.; Writing – Original Draft, S.S.S.; Writing – Review & Editing, S.S.S., J.S. and M.A.A.; Funding Acquisition, M.A.A.; Resources, M.A.A.; Supervision, M.A.A.

DECLERATION OF INTERESTS

The authors declare no competing interests.

FIGURE LEGENDS

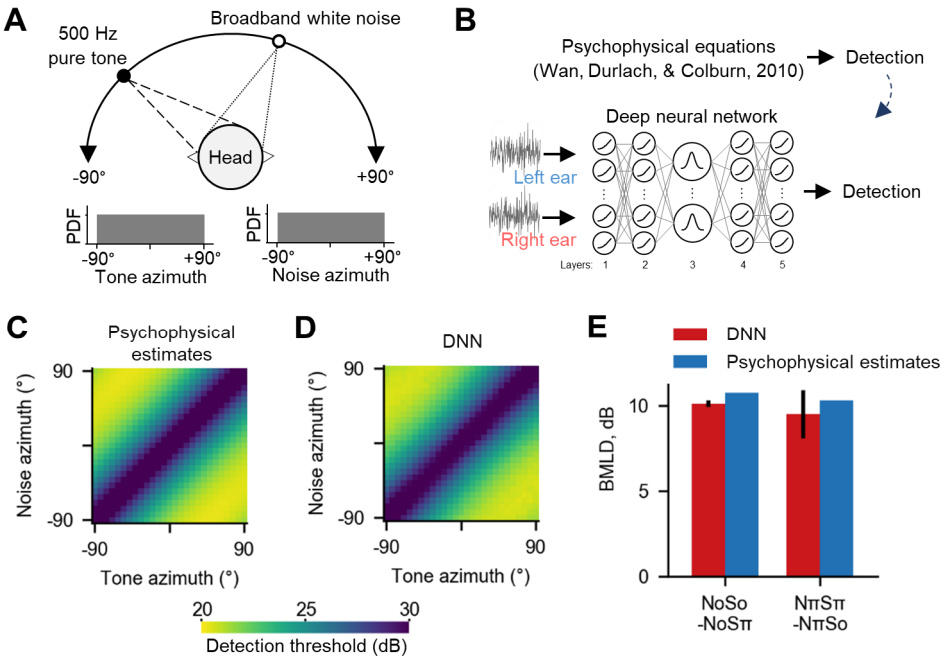

Figure 1. Deep neural network accounts for human binaural detection psychophysics.

(A) Data from a frontal field binaural detection task (generated using psychophysical equations) were used to train DNNs to detect a pure tone (black circle) in broadband noise (empty circle). Locations (and hence interaural time differences (ITDs)) of the tone and noise were chosen at random on each trial and were equally likely to come from each location (bottom panels).

(B) The DNN was a 5-layer network with a low-dimensional central layer, i.e. layer 3, designed to promote interpretation of the internal workings of the network.

(C) Psychophysical estimates of human detection thresholds for a tone in noise, each with ITDs mapped from differing azimuths.

(D) The DNN performance for unseen testing data configured as in (C).

(E) Binaural masking level differences (BMLDs) were derived for experimental stimulus configurations (NoSo/NoSπ, NπSπ/NπSo, Note: π is beyond the DNNs trained range). Error bars for DNN represent 95% confidence intervals.

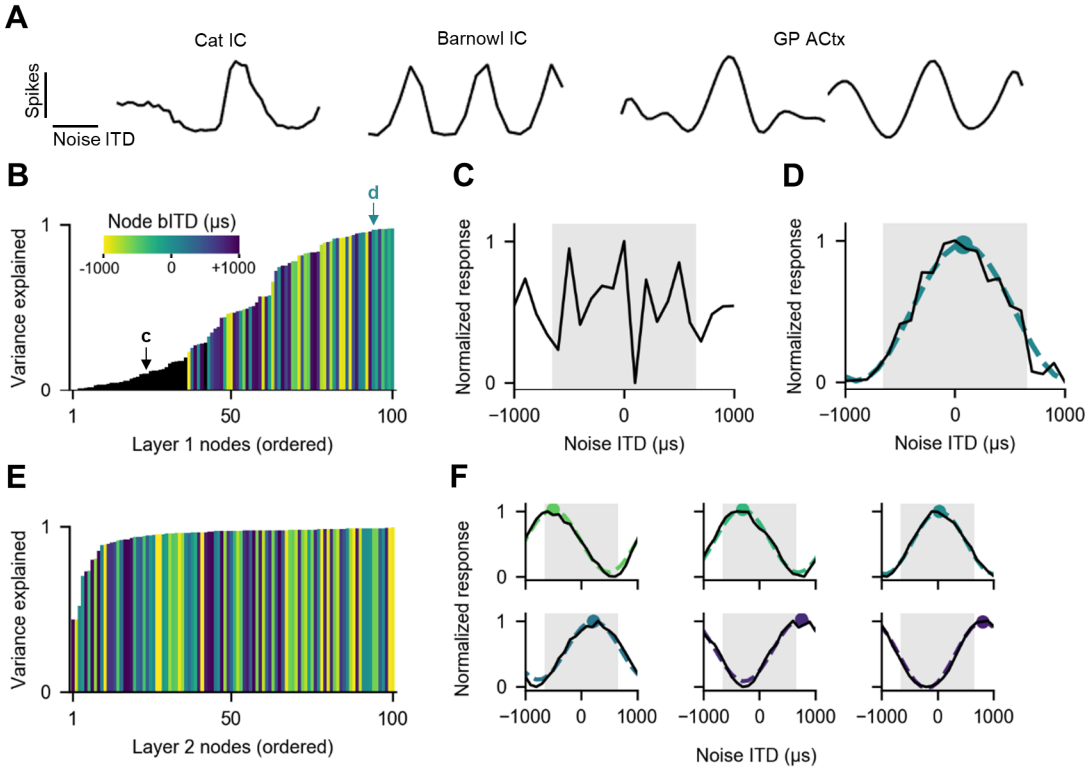

Figure 2. Time delay tuning emerges early in the network.

(A) Neural ITD tuning curves have been observed in a number of animal species, including cat[21] (inferior colliculus, IC), barnowl[25] (IC), and guinea pig[8] (auditory cortex, GP ACtx).

(B) ITD tuning emerged as a property of nodes within the layer 1 of the DNN. ITD tuning was defined as the proportion of variance explained (R^2) by fits (Gabor functions) regressed to noise delay responses for nodes in the DNN's layers. Bars are color-coded by the nodes' best ITD (bITD; black indicates the Gabor fit was not significantly better than a linear fit).

(C) Example of responses to noise ITDs in a layer 1 node with no clear tuning. The gray box underlays represent the ITD-limit for our simulation (modelled on the human head).

(D) Example of ITD responses in a layer 1 node tuned near 0 μs.

(E) ITD tuning was a dominant characteristic of nodes in layer 2.

(F) Individual examples of ITD tuning within a subselection of nodes in layer 2.

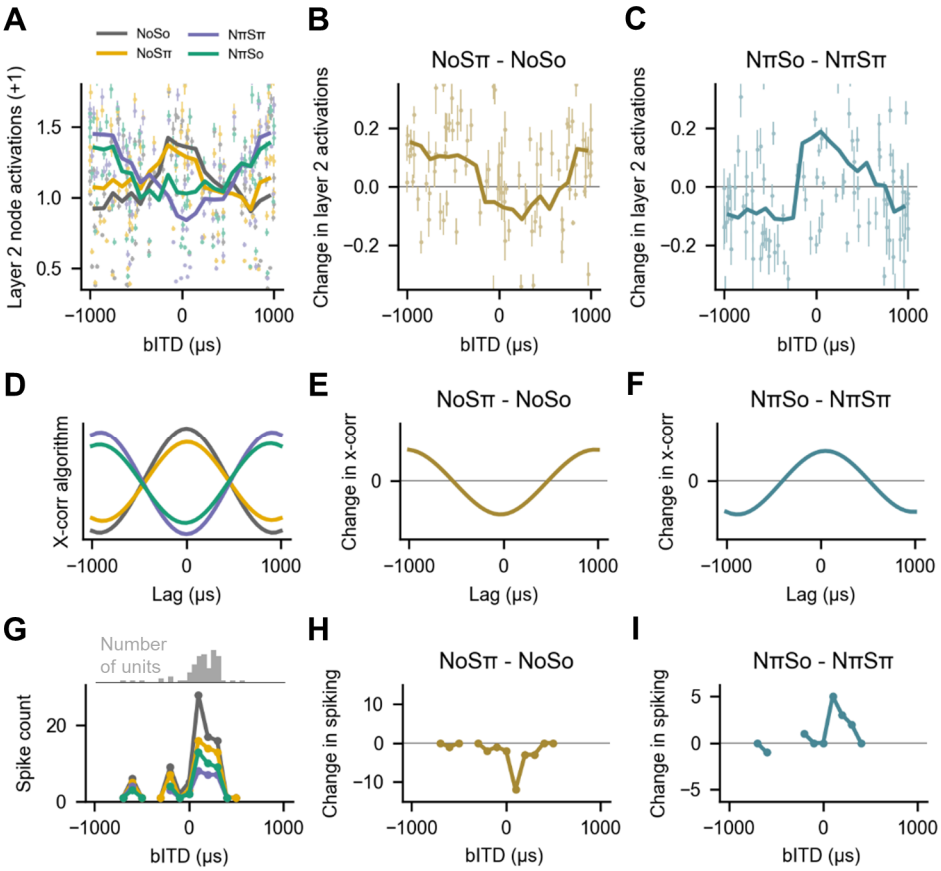

Figure 3. Network dynamics match those of a cross-correlation mechanism of the type suggested in animal physiology.

(A) The internal mechanism of layer 2 of the DNN was probed by considering activation of nodes with different ITD tuning for a set of typically employed binaural detection stimuli (NoSo, NoSπ, NπSπ, NπSo; color-coded). Single node data in light colors, where error bars represent 95% confidence intervals. Moving averages are overlaid and color-coded.

(B) The activation change of nodes to paired stimulus configuration, NoSo vs NoSπ, produced a profile across bITD (Error bars represent 95% confidence intervals. Moving averages are overlaid and color-coded).

(C) Equivalent to (B) for the stimulus configuration NπSπ vs NπSo.

(D,E,F) These profiles matched that of a simple cross-correlation algorithm.

(G,H,I) In addition, similar profiles have been observed in the guinea pig animal model (unit histogram is presented in (G)).

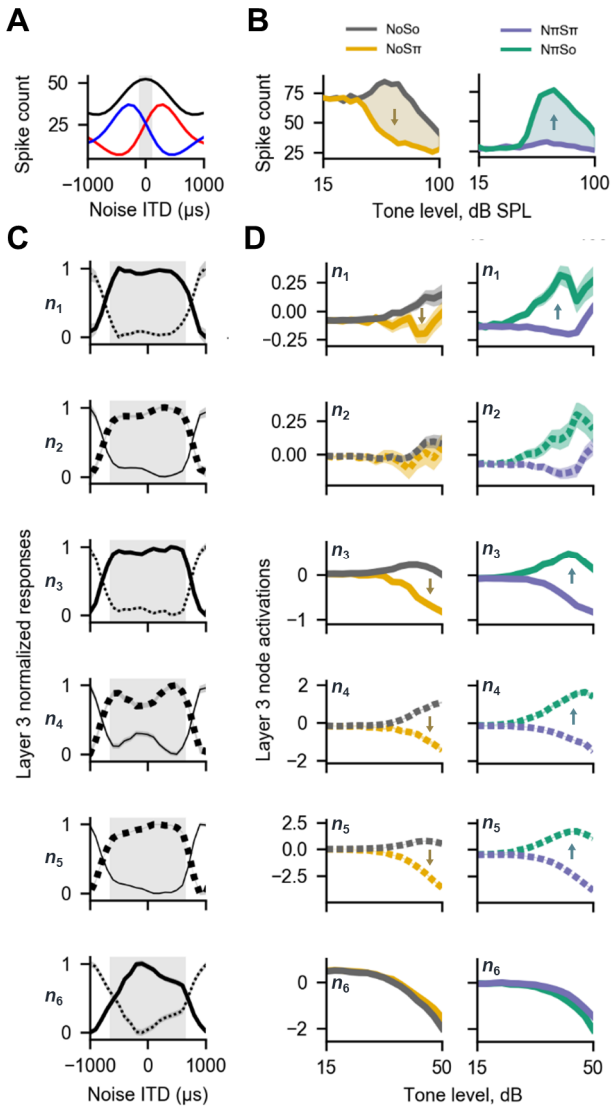

Figure 4. Low-dimensional network representations imitate neural signature of population-level cortical activity.

(A) A noise delay function from a representative neuron in right guinea pig auditory cortex[8] (red line) alongside its reflection representative of left cortex (blue line) and their sum (black line). The gray box underlays represent the ITD-limit for the guinea pig.

(B) Population masked rate-level functions recorded from guinea pig auditory cortex[8], in response to experimental binaural detection stimuli (NoSo, NoSπ, NπSπ, NπSo) are shown amongst arrows indicating changes as stimuli become more easily detectable.

(C) Noise delay functions from operational nodes in the DNN's layer 3 (n_1-n_6). The gray box underlays represent the ITD-limit for an average human head.

(D). Masked rate-level functions in response to the same stimuli configurations as in (B, i.e. NoSo, NoSπ, NπSπ, NπSo). Dashed lines represent polarity corrected responses (such that noise delay functions in (**C**) are peaked as in (**A**)). Lightly shaded regions represent 95% confidence intervals.

STAR ★ METHODS

RESOURCE AVAILABILITY

Lead Contact. Further information and requests for resources should be directed to the Lead Contact, Samuel Smith (samuel.smith@nottingham.ac.uk).

Materials Availability. This study did not generate any new unique materials.

Data and Code Availability. Code generated during this study is available at https://github.com/Hearing-Sciences/BinauralDetection_DNN.

METHOD DETAILS

Training stimuli. Stimuli parameters were selected to maximize comparative opportunities with published experimental data. Pure tones were produced at a frequency of 500 Hz, and presented at levels between 0 and 50 dB. Pure tones were 20 ms long (10 periods) and produced with a sample rate of 20 kHz. Pure tones were masked by randomly distributed broadband noise (50-5000 Hz, limited by 6th order Butterworth bandpass filter) with an overall level of 60 dB. The tone and noise were gated simultaneously. Horizontal perception of space at low frequencies is largely based upon ITDs[40]. Given this, tones and noises were simulated with ITDs mapped from two independent angles in the azimuth between -90° (far left) and 90° (far right). ITDs were derived from Woodworth's equation[19], assuming a head radius of 0.0875 m.

Binaural detection rates and thresholds. The theory of equalization and cancellation[12] has wide human psychophysical support, successful in predicting BMLD data[10,12] and binaural pitch phenomena[12,14,15], underpinning other models of binaural hearing[16], and proven psychophysically favourable relative to other prominent theories[17]. Detection thresholds were calculated from phenomenological equations derived from this theory[12,18]:

$$D(\tau_S, \tau_N) = 31 - 10\log_{10}\max\left\{\frac{k-\cos(\omega_0\tau_S-\omega_0\tau_N)}{k-\gamma(\tau_N-\tau_0)}, 1\right\}\text{ dB} \quad (1)$$

where τ_s and τ_N are the interaural time lags of the signal and noise, ω_0 is the angular frequency of the pure tone signal, $k = (1+\sigma_\varepsilon^2)e^{\omega_0^2\sigma_\delta^2}$ where σ_ε^2 and σ_δ^2 are jitter (internal noise) parameters, γ is the normalized envelope of the autocorrelation of the narrow-band noise output of a filter centred at the target tone frequency, and τ_0 is an optimal time equalization parameter. The parameters were chosen according to Durlach's[12] original formulation, e.g. γ assumes a filter with triangular gain characteristics. Psychometric functions

were derived under the assumption that detection thresholds represented a d' of 1 in a yes-no experiment[41,42]:

$$R(a, D) = \Phi\left(0.501611 \times 10^{0.1(a-D)}\right) \quad (2)$$

where a is pure tone amplitude and D the detection threshold (**Equation 1**).

Deep neural network. We trained DNNs[4,5] to predict the detection rates of tones presented amongst noise, with varying ITDs. Networks took 800 input values comprised of 400 samples from the left-ear waveform and 400 samples from the right-ear waveform. These inputs were passed through two 100 neuron exponential linear unit layers (ELU), referred to as the "decoder" portion of the network. This was followed by a layer of 10 latent Gaussian nodes ($\gg$ than the parameters varied in the generation of training stimuli) with minimal uncorrelated representations, constrained by a parameter β. This was followed by another two 100 neuron exponential linear unit layers, referred to as the "decision" portion of the network. All layers were fully connected and feed-forward. DNNs were trained and validated (95%/5% split respectively) on 10^6 instances of a random phase tone at a random level (0-50 dB) in randomly generated white noise, both presented with ITDs mapped from random angles in the azimuth, and the corresponding detection rates (**Equation 2**).

The Adam optimization algorithm was used to minimize the cost function:

$$C_\beta(\hat{x}, x, \sigma, \mu) = \|\hat{x} - x\|_2^2 - \frac{\beta}{2}\sum_i \log\left(\sigma_i^2\right) - \mu_i^2 - \sigma_i^2 \quad (3)$$

where $\hat{x}$ and x are predicted and true detection rates respectively, and σ and μ are the standard deviation and mean of latent Gaussian nodes respectively. Batch size (number of training instances employed in each iterative update of network parameters) was set to 256. The learning rate (training hyperparameter) was set to 5×10^{-4} for 1000 epochs (total passes of entire training dataset). Ten DNNs were trained for each value of β, namely 0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} (60 in total). The DNN with the least root mean square error, between predicted detection rates and ground truth, for the validation dataset, was selected for further analysis. Central nodes were considered operational if the Kullback–Leibler divergence between their individual responses and a unit Gaussian was larger than 0.1 bits (**Figure S1B**).

Network detection thresholds. For a given stimulus configuration, DNN detection thresholds were obtained by calculating the mean of 10 detection rates across tone levels set between 0 and 50 dB in 2.5 dB steps and regressing a psychometric curve (**Equation 2**). This was repeated 10 times for a given stimulus configuration. Stimuli for which detection thresholds were derived included:

- random phase tones amongst randomly generated broadband noise with ITDs each mapped from fixed azimuthal locations,
- and random phase tones and randomly generated broadband noise each either in or out of phase (i.e. NoSo, NoSπ, NπSπ, and NπSo).

Detection thresholds were also derived following ablations, where a set percentage of a given layers nodes were randomly nullified.

Node representations. Node activations were measured as a function of ITD for broadband noise (50-5000 Hz, 60 dB). ITDs ranged from -2000 μs to 2000 μs in steps of 100 μs. Node activations were also measured in response to So (in-phase) or Sπ (out-of-phase) signals masked by either No (in-phase) or Nπ (out-of-phase) broadband noise. Activations in layer 2 nodes were measured in response to a tone level of 35 dB amongst a noise level of 60 dB (in **Figure 3a**, activations were displayed with +1 added to their value because the minimum value of the ELU activation function is -1)[3]. In latent Gaussian nodes (in the central layer 3), masked rate-level functions to NoSo, NoSπ, NπSπ, and NπSo were measured amongst pure tone levels varied between 0 and 50 dB in 2.5 dB steps and 60 dB broadband noise. For all response measurements, stimuli were 20 ms long with a sample rate of 20 kHz. For a given stimulus configuration, activations were measured in response to 5000 random generations.

Binaural cross-correlation algorithm. For comparative purposes, outputs from a binaural cross-correlation algorithm were calculated[43]. The stimuli NoSo, NoSπ, NπSπ, and NπSo were generated for a 35 dB tone and a 60 dB randomly distributed broadband noise. Stimuli were sampled at 20 kHz and were 1 s in duration. Signals were passed through gammatone filters centered at 500 Hz and passed through a model of neural transduction[44]. The outputs were then delayed relative to one another, and the cross-products calculated and summated.

QUANTIFICATION AND STATISTICAL ANALYSIS

ITD tuning was quantified by fitting a Gabor function[22] to noise delay responses. The parametric expression for a Gabor function is:

$$G = Ae^{-(ITD-bITD)^2/2s^2}\cos\big(2\pi F(ITD-bITD)\big) + C \quad (4)$$

in which we characterized a nodes' best ITD as the parameter $bITD$, F is the tuning curve frequency, A is a scaling factor (constrained to be positive), C is a constant offset, and s is a decay constant. These parameters were fit with the non-linear least squares algorithm *curve_fit* (a SciPy function[45]). An F-test was used to assess whether a Gabor function was a significantly better fit to noise delay responses than a linear function of ITD.

We performed Student's two-tailed t-tests (assuming unequal variance) to assess BMLDs and differences in node responses at threshold tone levels. We also used Student's two-tailed t-tests (assuming unequal variance) to assess BMLDs following network ablations, for which a Bonferonni correction was applied to offset the impact of testing multiple ablation rates. Pearson product-moment correlation was calculated between the average responses of nodes to NoSo, NoSπ, NπSπ, and NπSo, and the delay matched outputs of a binaural

cross-correlation algorithm. A two-way ANOVA was also run for these nodes responses, with main effects of best ITD and noise phase. For the outlined statistical analyses, the criterion for significance (following multiple comparison corrections, when applied) was set to p=0.05. Error bars and lightly shaded underlays in figures are 95% confidence intervals.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Python 2.7	N/A	https://www.python.org/
Numpy	Oliphant, 2006	http://www.numpy.org/
Scipy	Jones et al., 2001	https://www.scipy.org/
TensorFlow	Abadi et al., 2016	https://www.tensorflow.org/
SciNET	Iten et al., 2018	https://github.com/eth-nn-physics/
MATLAB 2019b	Mathworks, MA, USA	https://www.mathworks.com/

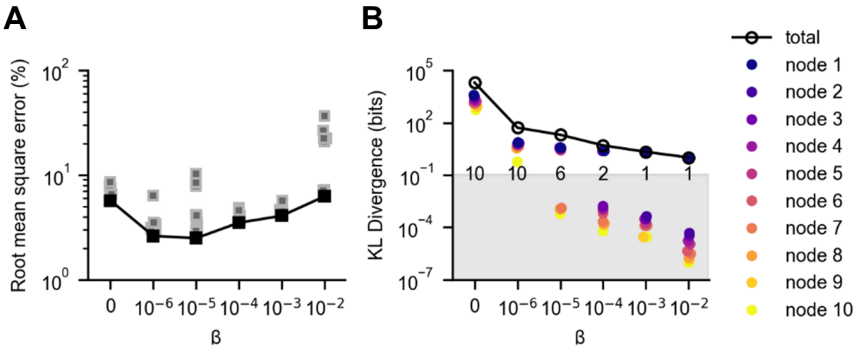

Figure S1. Meta-parameter search

(A) Prediction error for 60 DNNs (10 for each value of β, see Methods) tested on the validation dataset. The DNNs with the minimum error for each value of β are represented with large black squares.

(B) Information transmission was investigated for the most accurate DNNs for each value of β. Displayed is the total Kullback–Leibler (KL) divergence between latent nodes in layer 3 and an isotropic Gaussian (empty black circles). The KL divergence between each individual node and a unit Gaussian is also shown (color coded in order). The gray region represents nodes deemed to be suppressed during training. The number of nodes with KL divergences above this region (out of 10) are typed on the upper border of this region.

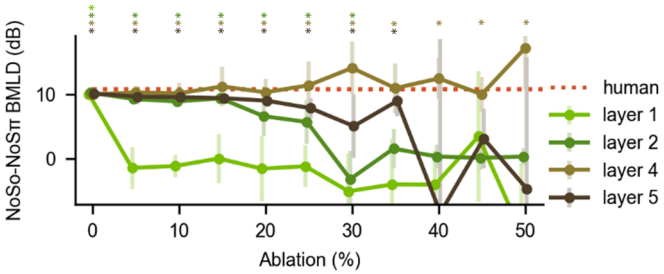

Figure S2. Early network ablation is detrimental to binaural detection.

BMLDs (NoSo-NoSπ) were predicted following varying levels of random ablation (setting to zero) to nodes in the DNN's layers. Error bars represent 95% confidence interval. Color-coded asterisks indicate significant BMLDs ($p<0.05$ following Bonferonni correction).

REFERENCES

1. Khatami, F., and Escabí, M.A. (2020). Spiking network optimized for word recognition in noise predicts auditory system hierarchy. PLOS Comput. Biol. *16*, e1007558.
2. Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S. V., and McDermott, J.H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron *98*, 630-644.e16.
3. Koumura, T., Terashima, H., and Furukawa, S. (2019). Cascaded tuning to amplitude modulation for natural sound recognition. J. Neurosci. *39*, 5517–5533.
4. Iten, R., Metger, T., Wilming, H., del Rio, L.L., and Renner, R. (2018). Discovering physical concepts with neural networks. Phys. Rev. Lett. *124*, 010508.
5. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). B-VAE: Learning basic visual concepts with a constrained variational framework. In 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.
6. Hirsh, I.J. (1948). The Influence of Interaural Phase on Interaural Summation and Inhibition. J. Acoust. Soc. Am. *20*, 536–544.
7. Hirsh, I.J., and Burgeat, M. (1958). Binaural Effects in Remote Masking. J. Acoust. Soc. Am. *30*, 827–832.
8. Gilbert, H.J., Shackleton, T.M., Krumbholz, K., and Palmer, A.R. (2015). The neural substrate for binaural masking level differences in the auditory cortex. J. Neurosci. *35*, 209–220.
9. Colburn, H.S. (1977). Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise. Cit. J. Acoust. Soc. Am. *61*, 525.
10. Domnitz, R.H., and Colburn, H.S. (1976). Analysis of binaural detection models for dependence on interaural target parameters. J. Acoust. Soc. Am. *59*, 598–601.
11. Jeffress, L.A. (1972). Binaural signal detection- Vector theory(Vector correlation theory and neural mechanisms of binaural signal detection in human auditory system). Found. Mod. Audit. theory. *2*, 351–368.
12. Durlach, N.I. (1963). Equalization and Cancellation Theory of Binaural Masking- Level Differences. J. Acoust. Soc. Am. *35*, 1206–1218.
13. Dietz, M., Lestang, J.H., Majdak, P., Stern, R.M., Marquardt, T., Ewert, S.D., Hartmann, W.M., and Goodman, D.F.M. (2018). A framework for testing and comparing binaural models. Hear. Res. *360*, 92–106.
14. Klein, M.A., and Hartmann, W.M. (1981). Binaural edge pitch. J. Acoust. Soc. Am. *70*, 51–61.
15. Hartmann, W.M., and McMillon, C.D. (2001). Binaural coherence edge pitch. J. Acoust. Soc. Am. *109*, 294–305.

16. Breebaart, J., van de Par, S., and Kohlrausch, A. (2001). Binaural processing model based on contralateral inhibition. I. Model structure. J. Acoust. Soc. Am. *110*, 1074–1088.
17. Culling, J.F. (2007). Evidence specifically favoring the equalization-cancellation theory of binaural unmasking. J. Acoust. Soc. Am. *122*, 2803.
18. Wan, R., Durlach, N.I., and Colburn, H.S. (2010). Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. J. Acoust. Soc. Am. *128*, 3678–3690.
19. Woodworth, R., Barber, B., and Schlosberg, H. (1954). Experimental psychology.
20. Durlach, N.I., and Colburn, H.S. (1978). Binaural phenomena. In Handbook of perception, Vol IV, Hearing, E. C. Carterette and M. P. Friedman, eds. (New York: Academic), pp. 365–447.
21. Joris, P., and Yin, T.C.T. (2007). A matter of time: internal delays in binaural processing. Trends Neurosci. *30*, 70–78.
22. Lane, C.C., and Delgutte, B. (2005). Neural correlates and mechanisms of spatial release from masking: Single-unit and population responses in the inferior colliculus. J. Neurophysiol. *94*, 1180–1198.
23. Singheiser, M., Gutfreund, Y., and Wagner, H. (2012). The representation of sound localization cues in the barn owl's inferior colliculus. Front. Neural Circuits *6*, 45.
24. McAlpine, D., Jiang, D., and Palmer, A.R. (1996). Binaural masking level differences in the inferior colliculus of the guinea pig. J. Acoust. Soc. Am. *100*, 490–503.
25. Asadollahi, A., Endler, F., Nelken, I., and Wagner, H. (2010). Neural correlates of binaural masking level difference in the inferior colliculus of the barn owl (Tyto alba). Eur. J. Neurosci. *32*, 606–618.
26. Jeffress, L.A. (1948). A place theory of sound localization. J. Comp. Physiol. Psychol. *41*, 35–39.
27. Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2018). Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. IEEE J. Sel. Top. Signal Process. *13*, 34–48.
28. Vecchiotti, P., Ma, N., Squartini, S., and Brown, G.J. (2019). End-to-end Binaural Sound Localisation from the Raw Waveform. In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (Institute of Electrical and Electronics Engineers Inc.), pp. 451–455.
29. Francl, A., and McDermott, J. (2020). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. bioRxiv, 2020.07.21.214486.
30. Malhotra, G., Evans, B.D., and Bowers, J.S. (2020). Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. Vision Res. *174*, 57–68.

31. Scholte, H.S. (2018). Fantastic DNimals and where to find them. Neuroimage *180*, 112–113.
32. Goodman, D.F.M., and Brette, R. (2010). Learning to localise sounds with spiking neural networks. In Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010.
33. Schnupp, J.W.H., and Carr, C.E. (2009). On hearing with more than one ear: Lessons from evolution. Nat. Neurosci. *12*, 692–697.
34. Ching, T.Y.C., van Wanrooy, E., Dillon, H., and Carter, L. (2011). Spatial release from masking in normal-hearing children and children who use hearing aids. J. Acoust. Soc. Am. *129*, 368–375.
35. Wack, D.S., Cox, J.L., Schirda, C. V., Magnano, C.R., Sussman, J.E., Henderson, D., and Burkard, R.F. (2012). Functional Anatomy of the Masking Level Difference, an fMRI Study. PLoS One *7*, e41263.
36. Wack, D.S., Polak, P., Furuyama, J., and Burkard, R.F. (2014). Masking Level Differences – A Diffusion Tensor Imaging and Functional MRI Study. PLoS One *9*, e88466.
37. Sasaki, T., Kawase, T., Nakasato, N., Kanno, A., Ogura, M., Tominaga, T., and Kobayashi, T. (2005). Neuromagnetic evaluation of binaural unmasking. Neuroimage *25*, 684–689.
38. Fowler, C.G. (2017). Electrophysiological evidence for the sources of the masking level difference. J. Speech, Lang. Hear. Res. *60*, 2364–2374.
39. Jiang, D., McAlpine, D., and Palmer, A.R. (1997). Detectability index measures of binaural masking level difference across populations of inferior colliculus neurons. J. Neurosci. *17*, 9331–9339.
40. Rayleigh, Lord (1907). XII. On our perception of sound direction . London, Edinburgh, Dublin Philos. Mag. J. Sci. *13*, 214–232.
41. Egan, J.P., Lindner, W.A., and Mcfadden4, D. (1969). Masking-level differences and the form of the psychometric functton'.
42. Ingleby, J.D. (1967). Signal detection theory and psychophysics. J. Sound Vib. *5*, 519–521.
43. Akeroyd, M. (2017). A binaural cross-correlogram toolbox for MATLAB.
44. Meddis, R., Hewitt, M.J., and Shackleton, T.M. (1990). Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse. J. Acoust. Soc. Am. *87*, 1813–1816.
45. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., and Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.