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

SUMMARY

Neural activity from animals is often used as a proxy for the human brain. However, due to distinct environmental pressures, the relevance of perceptual systems described in animal models can be unclear. This problem is accentuated when animal physiology and human behaviour are not in complete agreement, as is the case for binaural hearing-in-noise. As a means to bridge this gap we reverse-engineered artificial neural networks from binaural psychophysics. By comparing *in silico* "physiology" in neural networks with *in vivo* animal data, we were able to make inferences as to the basis of binaural perception in humans. We observed the emergence of highly specialized solutions to account for low frequency sound detection. Artificial neurons developed a sensitivity to temporal delays that increased hierarchically and were widely distributed in preference. Network dynamics were consistent with a cross-correlator, comparable to the type reported in animal physiology. Our results attest to the likely prominence of this neural mechanism in human biology. Moreover, this is a primary demonstration that deep learning can infer tangible neural mechanisms underlying auditory perception.

INTRODUCTION

Deep neural networks (DNNs) have been used to solve problems in many fields of research and are increasingly proving their worth in the field of neuroscience[1]. Recent DNN studies have effectively addressed questions of *why* the auditory system is organized the way it is (typically in the context of task optimality)[2–5]. However, when human auditory neurophysiology is itself ambiguous, or unknown, we must first question *what* it is, i.e. discover its underlying dynamics. With a few design changes, could DNNs be better leveraged to learn about the underlying human neurophysiology driving audition? We tested this idea by training a DNN configured specifically to mimic human auditory behaviour and investigated what this might reveal about the underlying neural mechanism(s). One potential stumbling block in answering this question is the black-box nature of DNNs. However, new network architectures that put mechanistic interpretability at their forefront (as have shown promise in the field of physics[6,7]) could help overcome this limitation.

An ideal context in which to examine these inferential properties of DNNs is one where it is unclear whether non-human neurophysiology satisfactorily explain human audition. Binaural detection[8,9], where interaural differences enhance the detectability of one sound (a signal) amongst another (e.g. a background noise) by up to 15 dB, represents one such instance. There is ongoing debate as to which of a number of theoretical frameworks best relate to human binaural detection[10–14]. For example, animal neural data lend support to a

theory of binaural cross-correlation[15–17]. Whereas, human behaviour appears to be equally well, if not better, described by a noise equalization and subtraction scheme[12,18,19]. These discrepancies have not been resolved with human imaging data[20–23], for which resolution and response variability are key limitations[24]. Further, binaural detection is a highly specialised auditory function for which deficits have real-world consequences[25,26]. DNNs may offer the opportunity to bridge this gap between animal and human data, and as yet, the inner workings of DNNs constructed to handle binaural audio have scarcely been considered[27–29].

Here, we reverse-engineered DNNs that accounted for human-like behaviour in a binaural detection task. To best facilitate a mechanistic understanding, the DNN architecture was configured to decode inputs into low-dimensional latent representations from which decisions were based[6,7]. Not only did the DNN that best mimicked human behaviour learn to utilize binaural information, but it did so in a way strikingly similar to that described in animal physiology. The work attests to the prominence of binaural cross-correlation as a solution to signal detection at low frequencies, and its likely incidence in humans.

RESULTS

To augment the availability of data, we trained deep neural networks (DNNs) on data from a simulated binaural detection task. These data were generated by a set of equations recognized as effective in predicting human binaural psychophysics[12,18,19,30–32]. 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 (**Fig. 1a**). We found that the DNN configuration that optimally predicted unseen binaural detection data did so with a root mean square error (RMSE) of 2.5% (**Extended Data Fig. 1a**). The dynamics of this optimally performing DNN are the focus of this article (summary analytics across other DNN configurations are shown in **Extended Data Fig. 1**).

Deep neural network accounts for 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 (**Fig. 1c,d**). For example, in diotic noise (noise ITD = 0) where the tone came from the left, the 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 tested the network on stimulus configurations typically employed to study binaural detection. These include tones and noise configurations where they 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 (note that for a pure tone IPD is linearly proportional to ITD). 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[33]). 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[34] and with estimates from psychophysical equations (10.7 dB and 10.3 dB respectively; **Fig. 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 (**Fig. 2a**). We found that ITD tuning emerged hierarchically within the lower 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 physiology studies[35]. For nodes in the DNN's first layer, we observed significant ITD tuning in 63 out of 100 nodes (**Fig. 2b**). The noise delay responses exhibited in these nodes were well described by a Gabor function[15], the combination of a cosine windowed by a Gaussian (overlaid in **Fig. 2d**). By the DNN's second layer, significant ITD tuning had emerged in all 100 nodes (**Fig. 2e**). Estimates of each nodes' best ITD (bITD) were inferred from the Gabor fits (in order to account for nodes that were oscillatory in their noise delay responses, a form of phase-ambiguity noted in physiology[36], 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.

We went on to measure responses to stimuli commonly presented in physiology work to specifically probe binaural detection (i.e. NoSo, NoSπ, NπSπ, NπSo). 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, **Fig. 3a**). 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. The effects of tone phase on node dynamics were more subtle, although these dynamics were also in accordance with a nodes' tuning properties (**Fig. 3b,c**). 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.

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)[37]. 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 physiology studies have reported neural responses consistent with a cross-correlation mechanism[15,16,38,39] (**Fig. 3g-i**).

Early network ablation is detrimental to binaural detection.

The functional importance of early layers in the DNN (i.e. the "decoder" portion of the network) 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 (**Extended Data Fig. 2**). 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 representations in DNN 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[7], and our ambition was for something analogous for binaural detection. In the DNN, six central nodes were

deemed operational (**Extended Data Fig. 1b**), and we found noteworthy similarities between these nodes and population level cortical responses in the guinea pig[16] (comparable observations have also been reported in other animals[15,39] and in the guinea pig inferior colliculus[38]). The central nodes could take the value of any real number, positive or negative (a necessary limitation imposed by the network architecture). Although not essential for our main conclusions, node responses are presented polarity-corrected to best correspond with noise ITD tuning in cortex (**Fig. 4a,c**; guinea pig auditory cortex).

An interesting feature of binaural processing can be seen when comparing detection behaviour and physiological data. Similar improvements in behavioural performance can be attributed to completely different alterations in network dynamics. For example, BMLDs across NoSo-NoSπ conditions are similar to those found across NπSπ-NπSo conditions (**Fig. 1e**). However, guinea pig neural data[16] suggests different neural dynamics underlie these similar BMLDs[16]. In cortical recordings, population spike counts dropped amongst an No signal, as a pure tone went from So to Sπ (left panel in **Fig. 4b**). Conversely, amongst an Nπ signal, as a pure tone transitioned from Sπ to So, population spike counts increased (right panel in **Fig. 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 **Fig. 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 Fig. 4d**). One latent node was qualitatively different to the others (node 6 of **Fig. 4d**), and seemed to encode offsets related to the interaural phase of noise (No v. Npi, $p\ll0.0001$, two-sided unpaired t-test).

DISCUSSION

We set out to discover the efficacy of DNNs as a means of exploring the underlying mechanisms involved in hearing, specifically binaural detection. To do this, we trained DNNs to exhibit binaural detection resembling human behaviour and then examined their internal dynamics as model organisms[40]. The application of DNNs in this way is a promising method in systems neuroscience[1]. However, the capacity for DNNs to offer mechanistic understanding beyond broader analogies with auditory processing[2–5] has yet to be established. This work demonstrated not only a number of key similarities with non-human binaural systems but, critically, the method implies that the human auditory system may use alike mechanisms.

Perhaps easy to overlook, the DNN was able to successfully utilize binaural discrepancies in auditory stimuli, as opposed to seeking an alternative strategy[41] and/or failing

to exhibit binaural detection behaviour. ITD tuning, a well-known characteristic of binaural neurons and normally described in the context of sound localization[35], emerged early in the DNN. Although this is a notable finding when considering the potential of DNN models, ITD tuning is axiomatic in most explanations of binaural detection[10]. This ambiguity was better clarified by second layer nodes whose dynamics, in response to tones presented in broadband noise, resembled a binaural cross-correlation mechanism[13]. This mechanism was not hard-coded into the network, but inferred. Latent nodes, central within the DNN's architecture, were also compatible with the downstream dynamics of a cross-correlation mechanism and resembled guinea pig cortical neural recordings[16]. These results help reinforce and unite findings supportive of binaural cross-correlation as the mechanism underlying binaural detection in people, as opposed to other explanations[12,18,19].

We also experimented with a technique analogous to neural ablation[42], observing knock-on effects to detection performance, finding that manipulations early in the system were most detrimental. This result is consistent with atypically small BMLDs associated with peripheral tumors in the human auditory system, but not central lesions[43]. The potential insights attainable with other experimental DNN techniques is an exciting prospect (e.g. techniques analogous to neural stimulation[6] in tandem with optogenetics data). Yet, comparison between DNNs and neural biology come accompanied by an asterisk. We make no claims of creating a general-purpose implementation of the human binaural system. 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[2,44]) could have merit, but any impact on interpretability should be carefully weighed.

Conclusion

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

METHODS

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[45]. 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[33], 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[12,46] and binaural pitch phenomena[12,30,31], underpinning other models of binaural hearing[19], and proven psychophysically favourable relative to other prominent theories[18]. Detection thresholds were calculated from phenomenological equations derived from this theory[12,32]:

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

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 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 [47,48]:

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

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

Deep neural network. We trained DNNs[6,7] 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 (≫ 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 RMSE, 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 (**Extended Data Fig. 1b**).

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)[4]. 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[49]. 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[50]. The outputs were then delayed relative to one another, and the cross-products calculated and summated.

Statistical analysis. ITD tuning was quantified by fitting a Gabor function[15] to noise delay responses. The parametric expression for a Gabor function is:

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

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[51]). 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.

Resource availability. Code generated during this study is available at https://github.com/Hearing-Sciences/BinauralDetection_DNN. Further information and requests for resources should be directed to the Lead Contact, Samuel Smith (samuel.smith@nottingham.ac.uk).

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.

COMPETING INTERESTS

The authors declare no competing interests.

FIGURE LEGENDS

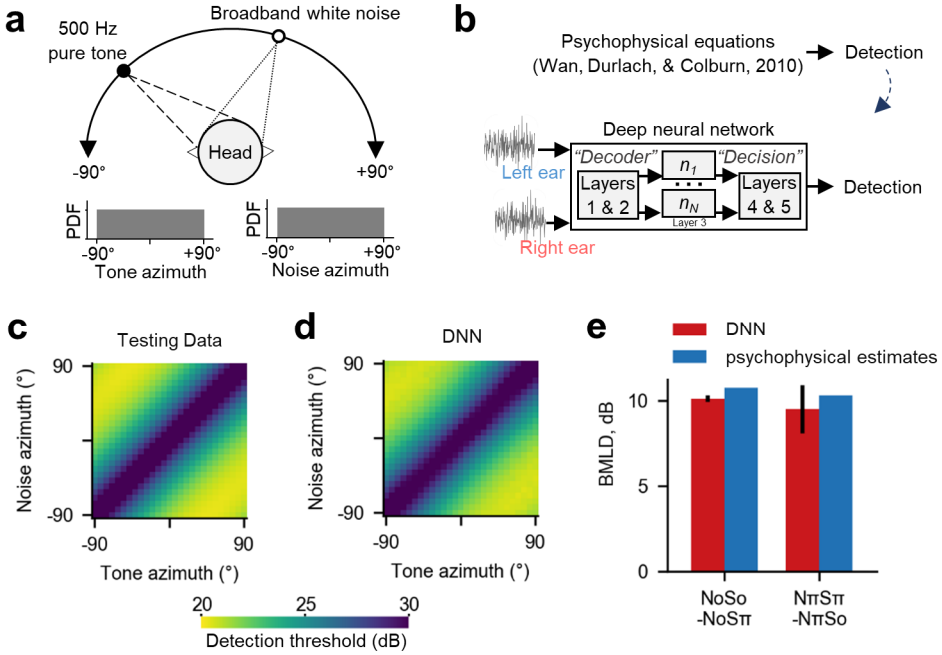
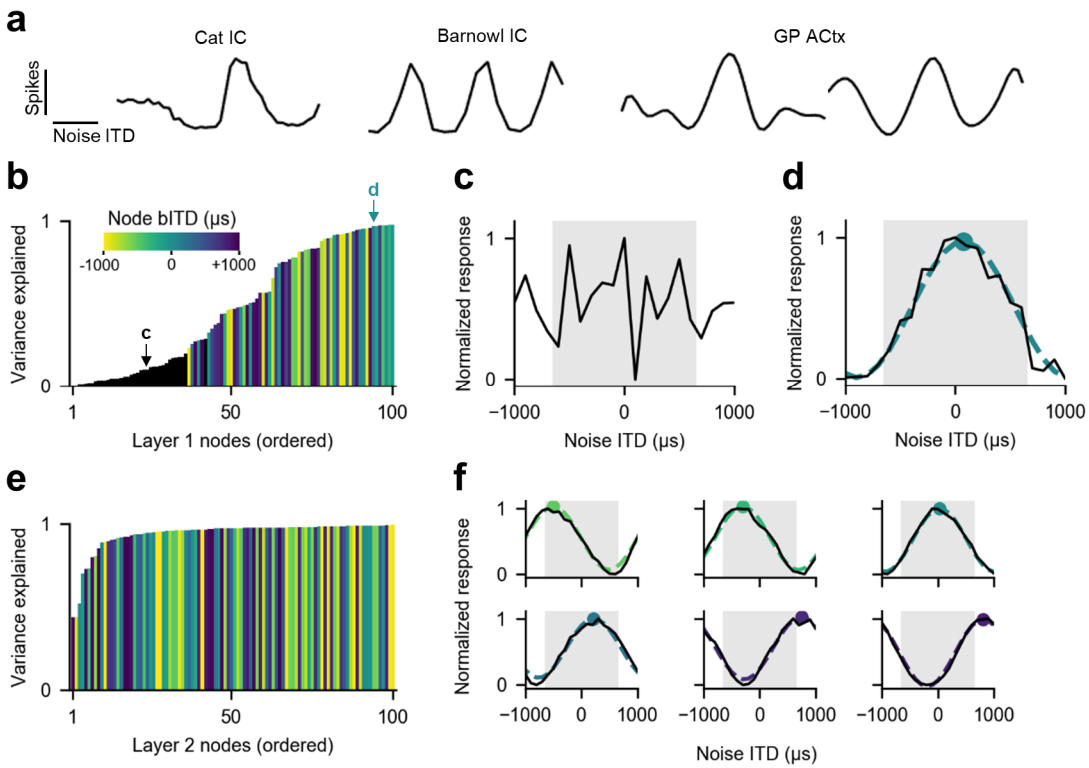


Figure 1. Deep neural network accounts for binaural detection psychophysics. Data from a frontal field binaural detection task (**a**, generated using psychophysical equations) were used to train DNNs (**b**) to detect a pure tone (black circle in **a**) in broadband noise (empty circle in a). Locations (and hence ITDs) of the tone and noise were chosen at random on each trial and were equally likely to come from each location (bottom panels of **a**). 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 (**b**). The DNN performance (**d**) for unseen testing data (**c**) was found to be comparable. In addition, binaural masking level differences (BMLDs) were derived for experimental stimulus configurations (NoSo/NoSπ, NπSπ/NπSo, **e**, Note: π is beyond the DNNs trained range). Error bars for DNN represent 95% confidence intervals.



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

375 Neural ITD tuning curves have been observed in a number of animal species (**a**), including
376 cat[35] (inferior colliculus, IC), barnowl[39] (IC), and guinea pig[16] (auditory cortex, GP ACtx). ITD
377 tuning emerged as a property of nodes within the early layers of the DNN and increased
378 hierarchically between layer 1 (**b**) and layer 2 (**e**). ITD tuning was defined as the proportion of
379 variance explained (R^2) by fits (Gabor functions) regressed to noise delay responses for nodes
380 in the DNN's layers. Bars are color-coded by the nodes' best ITD (bITD; black indicates the
381 Gabor fit was not significantly better than a linear fit). Individual examples of ITD tuning within
382 a subselection of nodes in layer 1 (**c** and **d**) and layer 2 (**f**). The gray box underlays represent
383 the ITD-limit for our simulation (modelled on the human head).

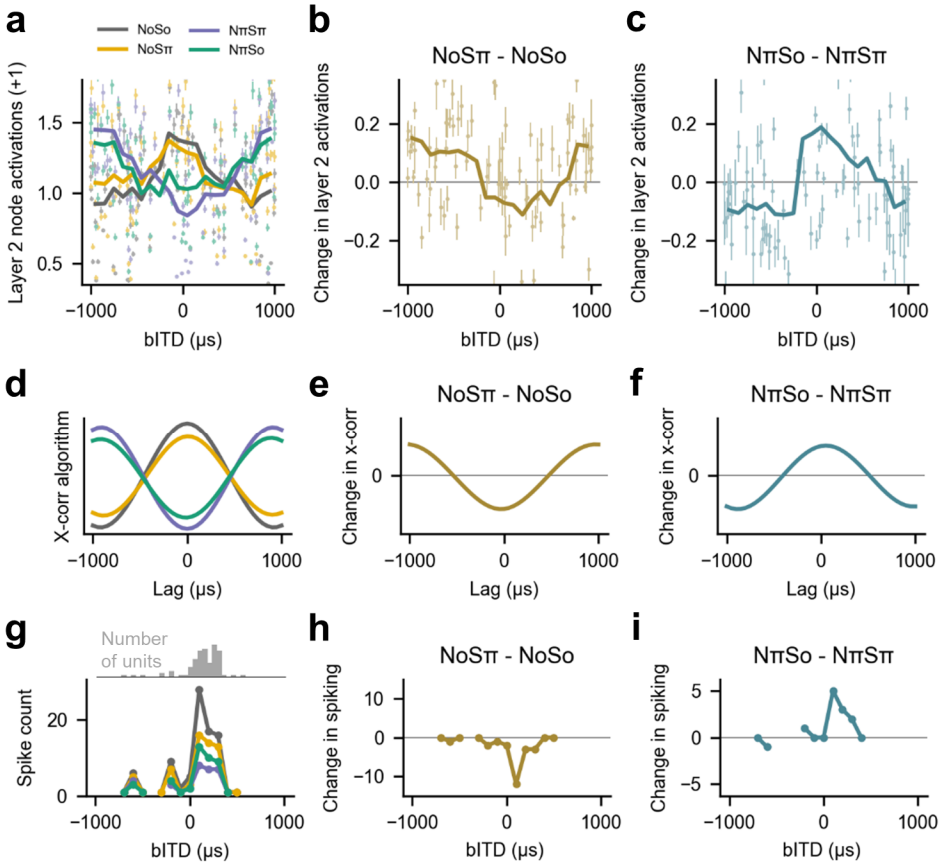


Figure 3. Network dynamics match those of a cross-correlation mechanism of the type suggested in animal physiology.

The internal mechanism of layer 2 of the DNN was probed by considering activation of nodes with different ITD tuning (**a**) 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. DNN activation (**a**) was found to closely match a simple cross-correlation model (**d**). It was also comparable to animal physiological data[16] (**g**) at the bITDs sampled (**g**, bottom panel shows the total spike counts of guinea pig auditory cortical neurons tuned to a given bITD, top panel shows neural count for each bITD). The activation change of nodes to paired stimulus configurations (NoSo vs NoSπ and NπSπ vs NπSo, **b** and **c** respectively) produced a profile across bITD (Error bars represent 95% confidence intervals. Moving averages are overlaid and color-coded). This profile matched that of a simple cross-correlation model (**e** and **f**, change in cross-correlation at different lags). In addition, a similar profile has been observed in the guinea pig animal model (**h** and **i**, change in spike count at different bITDs).

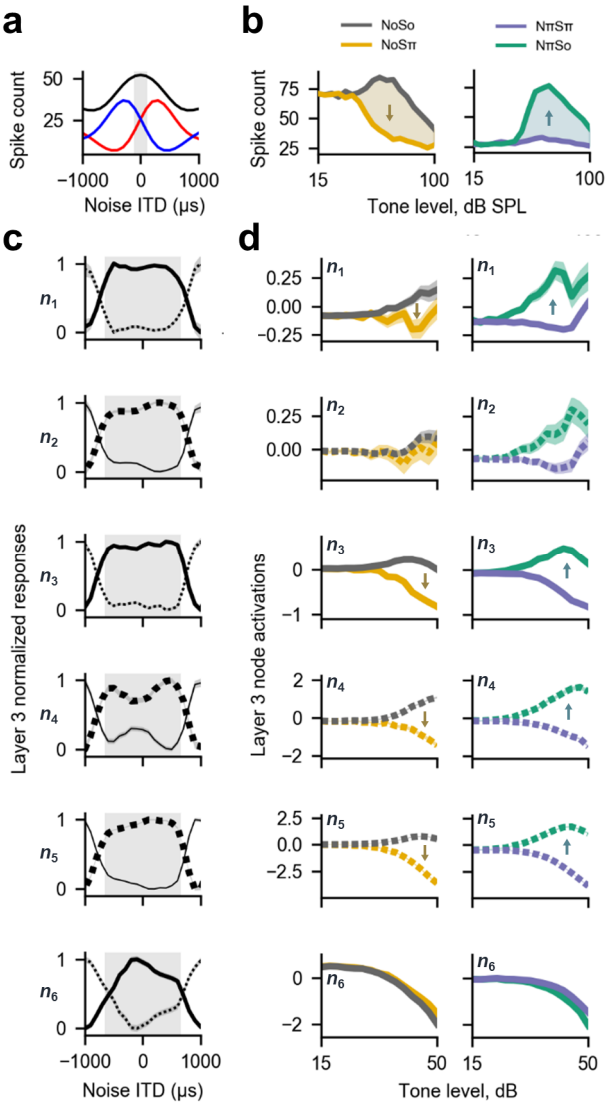
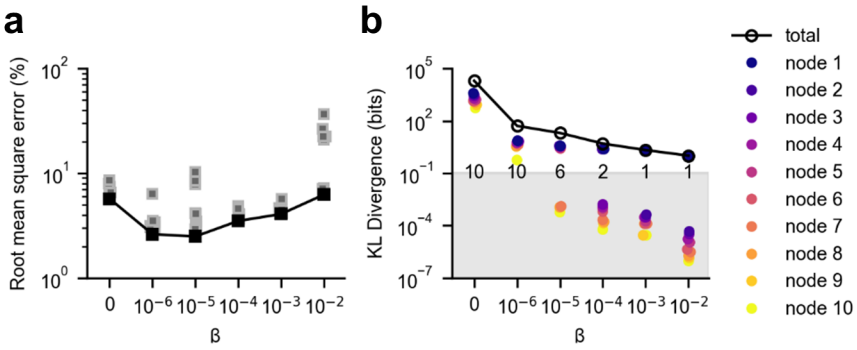


Figure 4. Low-dimensional representations in DNN imitate neural signature of population-level cortical activity.

A noise delay function from a representative neuron in right guinea pig auditory cortex[16] (**a**, red line) alongside its reflection representative of left cortex (**a**, blue line) and their sum (black line). Population masked rate-level functions recorded from guinea pig auditory cortex[16], in response to experimental binaural detection stimuli (NoSo, NoSπ, NπSπ, NπSo) are shown amongst arrows indicating changes as stimuli become more easily detectable (**b**). Noise delay functions (**c**) from operational nodes in layer 3 (n_1-n_6) are alongside masked rate-level functions (**d**) 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**)). The gray box underlays represent the ITD-limit for the guinea pig (in **a**) or an average human head (in **c**). Lightly shaded regions represent 95% confidence intervals.

412 # EXTENDED DATA



413 **Figure 1. Meta-parameter search**

414 Prediction error for 60 (10 for each value of β [see Methods]) DNNs tested on the validation
415 dataset (**a**). The DNNs with the minimum error for each value of β are represented with large
416 black squares. Information transmission was investigated for the most accurate DNNs for each
417 value of β (**b**). Displayed is the total Kullback–Leibler (KL) divergence between latent (layer 3)
418 nodes in layer 3 and an isotropic Gaussian (empty black circles). The KL divergence between
419 each individual node and a unit Gaussian is also shown (color coded in order). The gray region
420 represents nodes deemed to be suppressed during training. The number of nodes with KL
421 divergences above this region (out of 10) are typed on the upper border of this region.

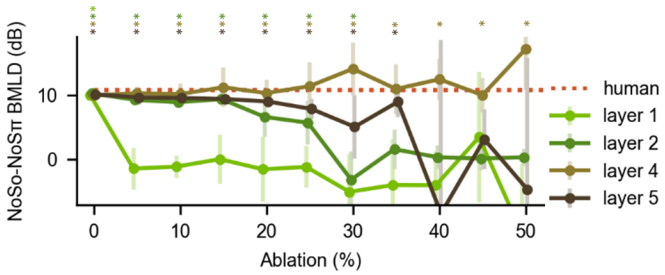


Figure 2. Early network ablation is detrimental to binaural detection.

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. Cichy, R. M. & Kaiser, D. Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences* vol. 23 305–317 (2019).
2. Khatami, F. & Escabí, M. A. Spiking network optimized for word recognition in noise predicts auditory system hierarchy. *PLOS Comput. Biol.* **16**, e1007558 (2020).
3. Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* **98**, 630-644.e16 (2018).
4. Koumura, T., Terashima, H. & Furukawa, S. Cascaded tuning to amplitude modulation for natural sound recognition. *J. Neurosci.* **39**, 5517–5533 (2019).
5. Furukawa, S., Terashima, H., Koumura, T. & Tsukano, H. Data-driven approaches for unveiling the neurophysiological functions of the auditory system. *Acoust. Sci. Technol.* **41**, 63–66 (2020).
6. Higgins, I. *et al.* B-VAE: Learning basic visual concepts with a constrained variational framework. in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (2017).
7. Iten, R., Metger, T., Wilming, H., del Rio, L. L. & Renner, R. Discovering physical concepts with neural networks. *Phys. Rev. Lett.* **124**, 010508 (2018).
8. Hish, I. J. Binaural summation and interaural inhibition as a function of the level of masking noise. *Am. J. Psychol.* **61**, 205–213 (1948).
9. Hirsh, I. J. The Influence of Interaural Phase on Interaural Summation and Inhibition. *J. Acoust. Soc. Am.* **20**, 536–544 (1948).
10. Domnitz, R. H. & Colburn, H. S. Analysis of binaural detection models for dependence on interaural target parameters. *J. Acoust. Soc. Am.* **59**, 598–601 (1976).
11. Jeffress, L. A. 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 (1972).
12. Durlach, N. I. Equalization and Cancellation Theory of Binaural Masking-Level Differences. *J. Acoust. Soc. Am.* **35**, 1206–1218 (1963).
13. Colburn, H. S. Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise. *Cit. J. Acoust. Soc. Am.* **61**, 525 (1977).
14. Dietz, M. *et al.* A framework for testing and comparing binaural models. *Hear. Res.* **360**, 92–106 (2018).
15. Lane, C. C. & Delgutte, B. Neural correlates and mechanisms of spatial release from masking: Single-unit and population responses in the inferior colliculus. *J. Neurophysiol.* **94**, 1180–1198 (2005).
16. Gilbert, H. J., Shackleton, T. M., Krumbholz, K. & Palmer, A. R. The neural substrate

for binaural masking level differences in the auditory cortex. *J. Neurosci.* **35**, 209–220 (2015).

17. Palmer, A. R. & Shackleton, T. M. The physiological basis of the binaural masking level difference. *Acta Acustica united with Acustica* vol. 88 312–319 (2002).
18. Culling, J. F. Evidence specifically favoring the equalization-cancellation theory of binaural unmasking. *J. Acoust. Soc. Am.* **122**, 2803 (2007).
19. Breebaart, J., van de Par, S. & Kohlrausch, A. Binaural processing model based on contralateral inhibition. I. Model structure. *J. Acoust. Soc. Am.* **110**, 1074–1088 (2001).
20. Wack, D. S. *et al.* Functional Anatomy of the Masking Level Difference, an fMRI Study. *PLoS One* **7**, e41263 (2012).
21. Wack, D. S., Polak, P., Furuyama, J. & Burkard, R. F. Masking Level Differences – A Diffusion Tensor Imaging and Functional MRI Study. *PLoS One* **9**, e88466 (2014).
22. Sasaki, T. *et al.* Neuromagnetic evaluation of binaural unmasking. *Neuroimage* **25**, 684–689 (2005).
23. Fowler, C. G. Electrophysiological evidence for the sources of the masking level difference. *Journal of Speech, Language, and Hearing Research* vol. 60 2364–2374 (2017).
24. Jiang, D., McAlpine, D. & Palmer, A. R. Detectability index measures of binaural masking level difference across populations of inferior colliculus neurons. *J. Neurosci.* **17**, 9331–9339 (1997).
25. Schnupp, J. W. H. & Carr, C. E. On hearing with more than one ear: Lessons from evolution. *Nature Neuroscience* vol. 12 692–697 (2009).
26. Ching, T. Y. C., van Wanrooy, E., Dillon, H. & Carter, L. Spatial release from masking in normal-hearing children and children who use hearing aids. *J. Acoust. Soc. Am.* **129**, 368–375 (2011).
27. Adavanne, S., Politis, A., Nikunen, J. & Virtanen, T. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. *IEEE J. Sel. Top. Signal Process.* **13**, 34–48 (2018).
28. Vecchiotti, P., Ma, N., Squartini, S. & Brown, G. J. End-to-end Binaural Sound Localisation from the Raw Waveform. in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* vols 2019-May 451–455 (Institute of Electrical and Electronics Engineers Inc., 2019).
29. Francl, A. & McDermott, J. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *bioRxiv* 2020.07.21.214486 (2020) doi:10.1101/2020.07.21.214486.
30. Klein, M. A. & Hartmann, W. M. Binaural edge pitch. *J. Acoust. Soc. Am.* **70**, 51–61 (1981).
31. Hartmann, W. M. & McMillon, C. D. Binaural coherence edge pitch. *J. Acoust. Soc. Am.* **109**, 294–305 (2001).

32. Wan, R., Durlach, N. I. & Colburn, H. S. Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *J. Acoust. Soc. Am.* **128**, 3678–3690 (2010).
33. Woodworth, R., Barber, B. & Schlosberg, H. Experimental psychology. (1954).
34. Durlach, N. I. & Colburn, H. S. Binaural phenomena. in *Handbook of perception, Vol IV, Hearing* (eds. Carterette, E. C. & Friedman, M. P.) 365–447 (New York: Academic, 1978).
35. Joris, P. & Yin, T. C. T. A matter of time: internal delays in binaural processing. *Trends in Neurosciences* vol. 30 70–78 (2007).
36. Singheiser, M., Gutfreund, Y. & Wagner, H. The representation of sound localization cues in the barn owl's inferior colliculus. *Front. Neural Circuits* **6**, 45 (2012).
37. Jeffress, L. A. A place theory of sound localization. *J. Comp. Physiol. Psychol.* **41**, 35–39 (1948).
38. McAlpine, D., Jiang, D. & Palmer, A. R. Binaural masking level differences in the inferior colliculus of the guinea pig. *J. Acoust. Soc. Am.* **100**, 490–503 (1996).
39. Asadollahi, A., Endler, F., Nelken, I. & Wagner, H. Neural correlates of binaural masking level difference in the inferior colliculus of the barn owl (Tyto alba). *Eur. J. Neurosci.* **32**, 606–618 (2010).
40. Scholte, H. S. Fantastic DNimals and where to find them. *NeuroImage* vol. 180 112–113 (2018).
41. Malhotra, G., Evans, B. D. & Bowers, J. S. Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Res.* **174**, 57–68 (2020).
42. Meyes, R., Lu, M., de Puiseau, C. W. & Meisen, T. *Ablation studies in artificial neural networks*. *arXiv* (arXiv, 2019).
43. Olsen, W. O., Noffsinger, D. & Carhart, R. Masking level differences encountered in clinical populations. *Int. J. Audiol.* **15**, 287–301 (1976).
44. Goodman, D. F. M. & Brette, R. 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* (2010).
45. Rayleigh, Lord. XII. On our perception of sound direction . *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **13**, 214–232 (1907).
46. Domnitz, R. H. & Colburn, H. S. Analysis of binaural detection models for dependence on interaural target parameters. *J. Acoust. Soc. Am.* **59**, 598–601 (1976).
47. Egan, J. P., Lindner, W. A. & Mcfadden4, D. *Masking-level differences and the form of the psychometric functton'*. (1969).
48. Ingleby, J. D. Signal detection theory and psychophysics. *J. Sound Vib.* **5**, 519–521 (1967).
49. Akeroyd, M. A binaural cross-correlogram toolbox for MATLAB. (2017).

50. Meddis, R., Hewitt, M. J. & Shackleton, T. M. Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse. *J. Acoust. Soc. Am.* **87**, 1813–1816 (1990).

51. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).