A maturational frequency discrimination deficit may explain developmental language disorder

Jones, Sam; Stewart, Hannah J; Westermann, Gert

Psychological Review

Published: 01/04/2024

Peer reviewed version

A maturational frequency discrimination deficit may explain developmental language disorder

Samuel David Jones[1, 2], Hannah Stewart[2], and Gert Westermann[2]

[1]Department of Psychology, Bangor University

[2]Department of Psychology, Lancaster University

13448 words

Author note

Correspondence concerning this article should be addressed to Samuel Jones, Room 309, Brigantia, Bangor University, Bangor, LL57 2AS. Email: samuel.jones@bangor.ac.uk.

Gert Westermann was supported by Economic and Social Research Council (ESRC) International Centre for Language and Communicative Development (LuCiD) [ES/S007113/1 and ES/L008955/1]. The theoretical view described in this manuscript was presented at the 28th Architectures and Mechanisms for Language Processing (AMLaP) conference. All data and materials required to re-run the simulations and analyses presented in this manuscript are available from the following public repository: https://osf.io/x2h8k/.

Abstract

Auditory perceptual deficits are widely observed among children with developmental language disorder (DLD). Yet the nature of these deficits and the extent to which they explain speech and language problems remain controversial. In this study, we hypothesise that disruption to the maturation of the basilar membrane may impede the optimisation of the auditory pathway from brainstem to cortex, curtailing high-resolution frequency sensitivity and the efficient spectral decomposition and encoding of natural speech. A series of computational simulations involving deep convolutional neural networks that were trained to encode, recognise, and retrieve naturalistic speech are presented to demonstrate the strength of this account. These neural networks were built on top of biologically truthful inner ear models developed to model human cochlea function, which – in the key innovation of the current study – were scheduled to mature at different rates over time. Delaying cochlea maturation qualitatively replicated the linguistic behaviour and neurophysiology of individuals with language learning difficulties in a number of ways, resulting in: (i) delayed language acquisition profiles; (ii) lower spoken word recognition accuracy; (iii) word finding and retrieval difficulties; (iv) 'fuzzy' and intersecting speech encodings and signatures of immature neural optimisation; and (v) emergent working memory and attentional deficits. These simulations illustrate the many negative cascading effects that a primary maturational frequency discrimination deficit may have on early language development, and generate precise and testable hypotheses for future research into the nature and cost of auditory processing deficits in children with language learning difficulties.

Keywords: developmental language disorder, auditory processing, spoken word recognition and retrieval, neural network, neural population geometry

A maturational frequency discrimination deficit may explain developmental language disorder

Introduction

There is astonishing variability in rates of early language development. Looking beyond population means, we see large windows of time in which language skills may emerge without any concern (Braginsky et al., 2018). Sometimes, however, a child's language is delayed enough to cause alarm among personal and professional caregivers. An estimated 7.5% of English-speaking children find acquiring and using language difficult enough to potentially interfere with their day-to-day emotional wellbeing and later with their educational outcomes (Norbury et al., 2016). Where such difficulties are evident in the absence of any obvious biomedical cause, such as Down's syndrome, the child may be diagnosed with developmental language disorder (DLD) and may undertake a tailored programme of language intervention targeting their specific areas of difficulty (Bishop et al., 2016).

Language disorder identification, assessment, and intervention are challenging because of the significant heterogeneity seen among affected children. Any aspect of language may be disrupted in DLD, from phonology through to syntax and pragmatics, and children often show concurrent developmental difficulties, for instance in motor control, or comorbidity with conditions such as developmental dyslexia or attention deficit hyperactivity disorder (ADHD) (Bishop et al., 2016). Furthermore, contrasting theoretical approaches have commonly centred on just one in a wide range of hypothesised cognitive faculties in accounting for discrete characteristics of this multifaceted profile. This approach has sometimes given the inaccurate impression that DLD is evidence of an isolated deficit in that faculty alone, for instance in working memory (Archibald & Gathercole, 2006), predictive

processing (Hestvik et al., 2022), lateral inhibition (McMurray et al., 2019), or statistical learning (Ullman & Pierpont, 2005).

The complex symptomology seen in DLD and overlap across associated diagnostic groups, at the level of both linguistic profile (i.e., from phonemes to pragmatics) and implicated cognitive faculties (e.g., working memory, statistical learning), has fostered a shift towards a 'transdiagnostic' mindset in neurodevelopmental disorder research (Astle et al., 2022). Here, focus is on what we might call *canonical* features of impairment – features sometimes termed 'bridging symptoms' – that hold widely not just within but often across diagnostic groups. Working memory deficits measured, for instance, in the nonword repetition and span tasks are widely considered one such canonical feature of developmental disorder, given that such deficits appear quite consistently across young children with a range of developmental difficulties (Archibald & Harder Griebeling, 2016; Gray et al., 2019; Henry & Botting, 2017).

Maintaining that there are canonical features of developmental disorder is, of course, very different from assuming there is a single cause of any given disorder. In general, contemporary research on early language disorder is averse to the notion that the varied profiles seen among children might have a single cause. This is perhaps a well-justified reaction to early research that held up DLD as evidence of an isolated deficit in an innately specified language acquisition device (a 'grammar module' of the brain encoded by the FOXP2 gene; Pinker, 1994) or similarly suggested that DLD was evidence of an discrete deficit in, for instance, working memory or statistical learning. We now know that the picture is considerably more complex. At the levels of genetics, neurobiology, and cognition, DLD appears to entail a constellation of causal mechanisms and risk factors (Bishop, 2006). A transdiagnostic, mechanism-centred approach fully appreciates this complexity and attempts to identify those dimensions of disorder that apply widely (though not *uniformly*) and which

may point us to better understanding and more effective intervention strategies (Fletcher-Watson, 2022). The careful, in-depth study of a specific and well-recognised canonical area of difficulty might show us how much we 'get for free' when we really explore the wide cascading effects implied by that area of difficulty.

The current study centres on one such canonical feature of developmental language disorder; auditory processing difficulties. While deficits in auditory perception are widely identified among children with neurodevelopmental disorder, most notably in DLD and dyslexia, the extent to which such deficits can explain early speech and language problems remains controversial (Bishop et al., 1999, 1999, 2012; Bishop & McArthur, 2005; Haake et al., 2013; McArthur & Bishop, 2004; Merzenich et al., 1996; Rosen, 2003; Tallal, 2013). In this study, we hypothesise that disruption to the maturation of the neural architecture underpinning high-resolution frequency discrimination from the prenatal period through the first two years of life (specifically, a disruption to basilar membrane maturation and resulting deficits in auditory brainstem optimization) may play a causal role in early speech and language disorder. Our account builds on prior work by McArthur and Bishop (2004) and Bishop and McArthur (2005), who first suggested that deficits in frequency discrimination may play an important role in the impairments observed among some children and adolescents with a diagnosis of DLD. In this study, we aim to substantially develop this account and to demonstrate its strength in a series of computational simulations that illustrate the varied consequences of a low-level frequency discrimination deficit within a controlled and transparent artificial learning environment. We aim to document the varied potential costs to early language development – i.e., the many cascading effects that we 'get for free' – as a result of a fundamental maturational deficit in frequency discrimination.

We begin this report by reviewing empirical research into the auditory processing skills of children with language disorder, highlighting an evolution from early theoretical

accounts centred around temporal processing, which relates to the speed at which the auditory system responds to acoustic input, to relatively recent accounts centred around frequency or *spectral* processing. We then review research into the maturation of the neural architecture supporting high-resolution frequency discrimination ability from the neonatal period through childhood, before considering how a disruption to this typical maturational trajectory might give rise to speech and language deficits. Subsequently, we present a computational model in which we simulate different rates of maturation in frequency discrimination ability while monitoring language acquisition rates, spoken word recognition accuracy, proxies for word finding latency, and neural speech representation integrity. We then discuss the implications of our results, the limitations of our computational approach, and directions for future investigation.

From temporal to spectral processing deficits in language disorder research

A dominant view developed principally through the work of Tallal and colleagues is that children with language learning difficulties have a primary deficit affecting the perception of acoustic signals that change rapidly, something that these authors refer to as a temporal processing deficit[1] (e.g., Merzenich et al., 1996; Tallal et al., 1981). Much of the empirical research in this direction made use of the auditory repetition task, or ART, in which children press buttons to identify changes in frequency in a series of pure tones. In the ART, performance accuracy among children with DLD was regularly shown to decrease significantly when inter-stimulus interval (ISI; i.e., the gap between tones) was reduced to below approximately 250 milliseconds, lending apparent support to the hypothesis that these children's auditory processing systems were ill-equipped to accurately perceive and encode rapidly unfolding natural speech (Merzenich et al., 1996; Tallal et al., 1981). This line of

[1] We note that the term 'temporal processing deficit' has been objected to on the basis that this body of research shows no evidence that the awareness of temporal order is compromised among children with DLD. The assumed difficulty instead relates to rapid changes in frequency, and so the term 'rapid perception deficit' may be more appropriate (Bishop, 2014, p. 90).

argument has been pursued in a significant body of research and has motivated the development of the Fast ForWord programme of intervention, which claims to be able to train sensitivity to rapidly occurring auditory stimuli through the controlled manipulation of ISI and in doing so confer gains in speech and language abilities (Tallal, 2013).

Despite the initial dominance of the temporal processing deficit hypothesis, however, a series of failed replications, both of the basic research and of the Fast ForWord intervention (Strong et al., 2011; Bishop & McArthur, 2005; McArthur & Bishop, 2004; see Rosen, 2003, for review) has motivated the search for alternative characterisations of the auditory perceptual deficits that appear to affect many children with speech and language problems. One promising, though comparatively underexamined view is that such deficits are spectral rather than temporal in nature (Bishop & McArthur, 2005; McArthur & Bishop, 2004; Mengler et al., 2005). That is, that these children's difficulty relates principally to distinguishing discrete sounds of similar frequency rather than discrete sounds that rapidly follow one another. For instance, across two studies Bishop and McArthur presented children aged 10 to 19 with and without language disorder with a baseline tone of 600Hz and a distinct tone which was initialised at 700Hz, but which was raised or lowered by increments of 25Hz to determine the minimal frequency discrimination threshold, or limen, that participants could identify (Bishop & McArthur, 2005; McArthur & Bishop, 2004; see also Mengler et al., 2005). These authors found that the minimal frequency discrimination threshold among children with severe language disorder was 750Hz (i.e., a 150Hz disparity) during an initial assessment and 674Hz at follow up (i.e., a 74Hz disparity), compared to 629Hz and 624Hz disparities respectively for control children. Readers may wish to visit one of the many freely available online pure tone generators to compare tones in this range themselves. For many, the average difference between the minimal threshold tones identified by children with DLD (i.e., 600Hz and 750Hz or 674Hz) will appear striking, attesting to the

difficulty such a deficit may cause during the analysis of the complex spectral profiles of natural speech (Nuttall et al., 2018; Sumner et al., 2018).

Crucially, Bishop and McArthur found that this deficit in frequency discrimination was observed regardless of the rate of stimulus presentation, providing compelling evidence that the auditory processing difficulties of some children affected by language disorder are spectral rather than temporal in nature, and perhaps explaining the failed replications of key studies in the temporal processing deficit literature (Bishop & McArthur, 2005; McArthur & Bishop, 2004; Mengler et al., 2005; Rosen, 2003; Strong et al., 2011). What is more, even those children with DLD who performed well in the behavioural tone discrimination task nevertheless showed immature waveforms during electroencephalography (EEG) monitoring, providing tentative support for the maturational account that Bishop and McArthur (2005) then offer to explain their findings.

The maturation of frequency discrimination skills

Bishop and McArthur (2005) explain their results in terms of a disruption to the typical maturation of high-resolution frequency discrimination. In order to situate this account, upon which we intend to elaborate, it is useful to review key research on the early maturation of frequency discrimination skills, and the neural basis of these skills. In younger children and infants, probing the maturation of frequency discrimination skills presents a significant challenge. Paradigms such as head turning and high-amplitude sucking have provided mixed results and are open to interpretation, not least that a failure to discriminate tones in such paradigms may be the result of immature motor skills or attention (see Burnham & Mattock, 2014, for review). In response, some researchers have advocated the use of neuroimaging methods such as EEG and magnetoencephalography when studying frequency discrimination in neonates and infants (e.g., Novitski et al., 2007). Despite their own

limitations, such neuroimaging methods are often considered to provide an index of neural activity that is relatively independent of motor and attentional factors (Novitski et al., 2007).

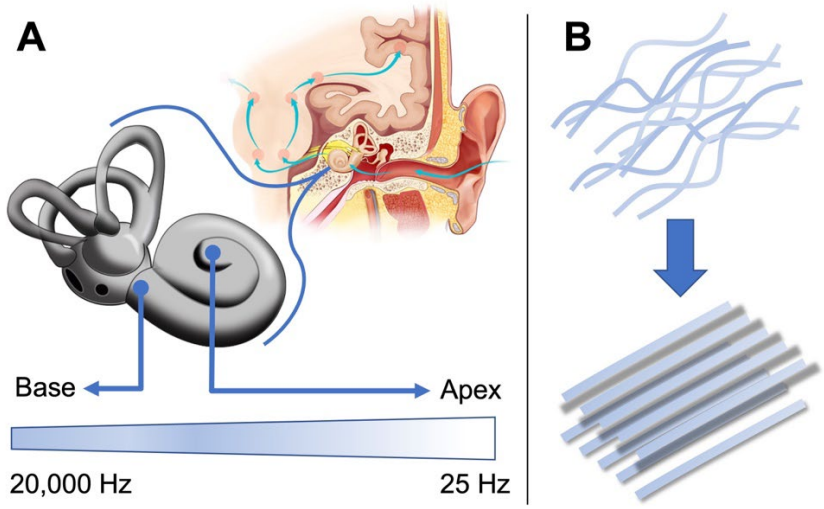
Neuroimaging involving neonates and infants corroborates indications from behavioural research of an early maturation in frequency discrimination ability (Jensen & Neff, 1993; Lopez-Poveda, 2014; Novitski et al., 2007; Shafer et al., 2000; Tharpe & Ashmead, 2001). This maturation is not uniform. High-frequency tone discrimination is approximately adult-like in apparently typically developing infants by six months of age. In contrast, low-frequency discrimination, in the range more regularly associated with speech signals (e.g., 400Hz), develops more slowly, with continued maturation apparent in children up to ages seven to nine (Jensen & Neff, 1993; Burnham & Mattock, 2014). While the empirical data vary somewhat, estimates from the 'odd one out' paradigm (also known as the 'mismatch negativity paradigm') suggest that newborns can detect a 20% though not a 5% change in frequency in a 250Hz-4000Hz window (Novitski et al., 2007; see Burnham & Mattock, 2014, for review). Such findings support the view that frequency resolution improves considerably from birth through childhood, making it increasingly easy to discriminate competing acoustic signals, and thus to perform the complex spectral analysis that accurate and efficient natural speech perception and encoding requires (Nuttall et al., 2018; Sumner et al., 2018).

The maturation of frequency discrimination skills reflects changes in neural architecture that, though many important questions remain, are now in a large part reasonably well understood. A key characteristic of the auditory perceptual system upon which speech representation and use is based is its tonotopic structure. That is, throughout the auditory pathway, from the inner ear to the auditory brainstem and on to the auditory cortex, we see selective responsivity to acoustic input of particular frequencies among sensory cells and neurons that constitutes the neural basis of frequency resolution and the decomposition of

auditory signals including speech (Echteler et al., 1989; Nuttall et al., 2018; Sumner et al., 2018). The characteristic 'tonotopic' structure of the auditory pathway results predominantly from the physical properties of the basilar membrane, a 35-mm coiled membrane within the inner ear (Figure 1A).

Figure 1

Schematic of Frequency Tuning and Structural Development in the Mammalian Cochlea



Note. Panel A shows the location of the cochlea in the inner ear (coloured inset), its coiled structure (in grey), and the mechanical frequency sensitivity gradient from base to apex of the basilar membrane within the cochlea. Panel B illustrates the development of basilar membrane micro-structure supporting high-resolution frequency tuning, from fibres that are low-diameter, sparse, and 'braided', to fibres that are higher-diameter, dense, and regular. The Panel A auditory system image (coloured inset) is in the public domain (https://commons.wikimedia.org/wiki/File:Hearing_mechanics.png). The Panel A greyscale cochlea image is available under a Creative Commons Attribution-Share Alike 4.0 International license (https://commons.wikimedia.org/wiki/File:Inner_ear.png).

The basilar membrane is narrow and firm at its base, and as a result of these physical properties fibres in this basal region vibrate maximally to the high frequencies in auditory

input (Figure 1A; Sumner et al., 2018). The apex of the basilar membrane is, in contrast, wide and relatively slack, and as a result fibres in this apical region vibrate maximally to the low frequencies in auditory input (Figure 1A; Sumner et al., 2018). For instance, voiceless fricatives such as /ʃ/, which contain relatively high-frequency components, may stimulate basal regions of the membrane, while vowels such as /ɑː/, which contain low-frequency components, may stimulate apical regions. Upon the basilar membrane sit a single row of approximately 3500 inner hair cells which become selectively responsive to specific frequencies – that is, they are 'frequency-tuned' – as a result of their position on the basilar membrane (Sumner et al., 2018; Tani et al., 2021). In turn, inner hair cells are innervated by spiral ganglion neurons which project to the cochlear nucleus, with this and subsequent innervation conserving tonotopic sensitivity and resulting in the emergence of frequency sensitive 'maps' throughout a complex array of subcortical structures of the auditory brainstem and on to the peripheral auditory cortex. The physical properties of the basilar membrane are, therefore, at the heart of frequency sensitivity and acoustic signal decomposition across the auditory pathway, and this itself underpins accurate and efficient speech processing and encoding (Burnham & Mattock, 2014; Echteler et al., 1989; Nuttall et al., 2018; Sumner et al., 2018; Tani et al., 2021). From the third trimester to 6 months of age structures from the auditory nerve throughout the auditory pathway to the auditory cortex undergo substantial changes in synaptic organisation, myelination, and dendritic arborisation, and this process of maturation continues through to two years of age during a typically rich period of language development (Chonchaiya et al., 2013). Work by Chonchaiya et al. (2013) indicates that, by nine months of age, auditory brainstem responses continuous with relatively mature brainstem organisation are predictive of better language outcomes.

Recent research has cast light on how the pre- and postnatal structural development of the basilar membrane underpins the emergence of high-resolution frequency tuning across the

auditory-linguistic pathway. Studies using electron microscopy and polarized light microscopy have shown that the basilar membrane is composed of collagenous filaments, or fibres, which are initially relatively low diameter, sparsely organised, and 'braided', but which increase in diameter, density, and linear regularity throughout early development (Figure 1B). Such studies have also determined an uneven time course in which structural maturation is slower in the membrane apex than it is in basal regions, a finding consistent with behavioural and neurophysiological evidence that low frequency component tuning comes online relatively slowly (Burnham & Mattock, 2014; Novitski et al., 2007; Tani et al., 2021). Animal models also provide mounting evidence that the protein coding gene *emilin 2* (elastin microfiber interfacer 2), which is part of the emilin family of glycoproteins that contribute in part to tissue elasticity, can seriously disrupt fibre development in the basilar membrane – i.e., can curtail typical increases in fibre diameter, density, and linear regularity – and can, therefore, disrupt the membrane's capacity to propagate frequency sensitivity throughout posterior structures of the auditory pathway supporting accurate and efficient frequency decomposition (Amma et al., 2003; Russell et al., 2020; Tani et al., 2021). This literature demonstrates how a genetic abnormality can in principle disrupt the emergence of the mechanical gradient of the basilar membrane.

Towards a maturational account of frequency resolution deficits and speech and language difficulties

Before stating our hypothesis, let us take stock of the key points reviewed so far:

1. Auditory processing deficits are widespread among children with DLD, and these deficits appear to be frequency-based rather than temporal in nature.
2. Evidence that deficits are related to frequency analysis points to specific cellular and neural structures of the auditory pathway. Specifically, the basilar membrane is at the heart of frequency tuning across the auditory pathway, with tonotopic maps emerging

throughout the auditory brainstem and cortex predominantly as a result of dynamic adaptation to the structural properties – i.e., the *mechanical gradient* – of the basilar membrane.

3. The basilar membrane undergoes crucial structural changes early in development, with the fibres from which the membrane is composed increasing in diameter, density, and regularity, in part as a result of *emilin 2* expression. This process of maturation is integral to the emergence of tonotopy across the auditory pathway.

Our hypothesis is, then, that:

> *Early disruption to the maturation of the physical properties of the basilar membrane which underpin that membrane's mechanical gradient (i.e., increases in fibre density, diameter, and linear regularity) may disturb the optimisation of the posterior auditory pathway from the brainstem to the cortex, curtailing high-resolution tonotopic sensitivity and contributing to speech and language difficulties in some children.*

The auditory pathway is, of course, a highly complex system, which could be disrupted by any number of influences operating across any number of its subsystems. It is, for instance, possible that auditory brainstem and auditory cortex optimisation are disrupted despite a properly maturing basilar membrane. A range of such alternative possibilities are presented in our *Discussion* section. Nevertheless, we believe that the hypothesis above provides a strong starting point for investigation given that (i) the auditory processing deficits we see in DLD appear to be spectral in nature and (ii) that a fully matured basilar membrane sits at the heart of high-resolution frequency processing across the auditory pathway. Our hope is that this literature review has shown that – though more work is undoubtedly required – there already exists a great deal of empirical evidence bearing on typical and atypical auditory pathway maturation and the potential impact of a maturational delay in this area on the emergence of speech and language. In our view, what is currently required to direct future investigation is a

compelling theoretical account linking these fragmentary research strands, and this is what we attempt to provide in the current study. Our aim is emphatically not to suggest that frequency discrimination deficits wholly explain early language disorder. Instead, we aim to flesh out one candidate mechanistic pathway within a complex constellation of many.

In what follows we simulate and monitor the dynamic adaptation of an artificial auditory-linguistic pathway (broadly auditory brainstem to cortex) in response to biologically plausible representations of speech-elicited activation patterns in the developing cochlea, under (i) non-developmental, (ii) regular, and (iii) delayed maturational trajectories. We show how a disruption to the maturation of cochlea microarchitecture may result in the atypical optimisation of subsequent neural pathways, qualitatively accounting for several commonly recorded characteristics of atypical human linguistic behaviour and neurophysiology, namely: (i) delayed language acquisition profiles (e.g., Norbury et al., 2016); (ii) spoken word recognition deficits (Andreu et al., 2012; Evans et al., 2018; Rispens et al., 2015; Velez & Schwartz, 2010); (iii) word finding or retrieval problems (Kambanaros et al., 2015; Messer & Dockrell, 2006); (iv) 'fuzzy' long-term speech representations (Claessen et al., 2009); (v) atypical neural signatures of auditory signal processing (e.g., Bishop & McArthur, 2005); and (vi) apparent working memory deficits, attributable, we argue, to the imprecision of activated long-term speech representations (Henry & Botting, 2017; Jones & Westermann, 2022).

Overview of simulations[2]

Network and training and testing regimes

The architecture used in these simulations is an artificial neural network known as a deep convolutional neural network. The work of McDermott and colleagues has been instrumental in demonstrating that despite obvious disparities between the biological auditory

[2] This paper is associated with a fully annotated Jupyter notebook (Kluyver et al., 2016), which is available from the following public repository and which can be used to replicate the simulations described or to experiment with alternative parameter configurations: https://osf.io/x2h8k/.

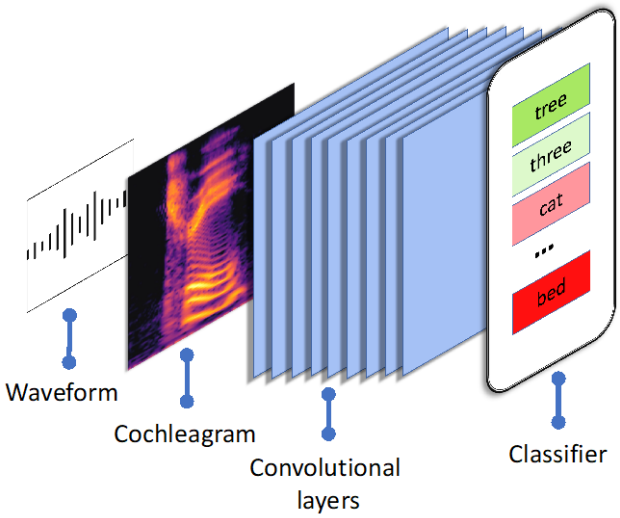
pathway and this artificial counterpart, including in general complexity and in learning procedures (see *Discussion*), close parallels are observed between convolutional neural network activity and human behavioural and neural responses across a wide range of tasks, such as speech localization, pitch perception, and hearing in noise (Francl & McDermott, 2022; Kell et al., 2018; Saddler et al., 2021). Convolutional neural networks are not 'circuit models' of the brain. That is, these networks are not intended to explicitly model fine-grained physiology such as ion channel behaviour (e.g., see Higgins et al., 2017, for a circuit model of speech perception and category formation). Rather, convolutional neural networks can provide high-order 'computational' insight, in the sense of Marr (1982), into how a perceptual processing hierarchy dynamically adapts to a particular form of input to solve a certain problem under varying constraints.

Our simulations made use of the ResNet-18 deep convolutional neural network (He et al., 2015), which we implemented using PyTorch (Paszke et al., 2019) in Python (Python Software Foundation, 2008). A full network description can be retrieved by running the Jupyter notebook associated with this project. Note that following the code examples associated with Stephenson et al. (2020; see https://github.com/schung039/neural_manifolds_replicaMFT), many of our analyses centre around the networks' 20 convolutional layers. For this reason, these layers are detailed in the Appendix alongside key hyperparameters. A total of nine convolutional neural networks (n=3; conditions defined below) were trained and tested on spoken words from the speech commands dataset (Warden, 2018), which contains 105,829 one-second spoken word waveforms of 35 word types (Figure 2). The speech commands dataset was chosen for this project because it is free and openly available, and because it is perhaps unique in comprising such a large number of exemplars of natural speech. Limitations of the speech commands dataset are noted in our discussion.

Over ten cycles, or ‘epochs’, of training, networks were required to categorise each spoken word that they perceived by outputting a probability distribution over their 35-word lexicon. The word with the highest probability assigned was taken as the networks’ selection. Networks responded dynamically to error signals propagated upon an incorrect classification by updating their inner weight matrices using the backpropagation algorithm after each spoken word exposure (i.e., batch size = 1) in order to reduce the future error rate. This constitutes a broad computational analogy to fluctuation in synaptic connection strength due to long-term potentiation (Lillicrap et al., 2020). Throughout training, networks were presented with random samples of 4000 exemplars per-epoch from the speech commands dataset. Random samples were matched within epochs across the network groups we define below. For instance, network one in each experimental condition saw the same random samples of training data, which differed in each training epoch. This ensures that any later-observed group-level performance discrepancies are not a function of differences in the data that the network has been trained on. We note that there is nothing special about the word as a unit of representation here. Our choice of dataset principally reflects its scale and the fact that it contains authentic spoken words, and similar effects would be expected were we modelling phonemes or multi-word constructions.

Figure 2

Neural Network Schematic



Note. Authentic, raw spoken waveforms are first passed through a cochlea model, before being passed through the deep convolutional neural network and the 35-way classifier.

Later, at test, neural networks were presented with another random sample of 1000 words from the speech commands dataset, a random sample which was again matched across conditions (defined below). We recorded a range of test performance metrics including speech recognition accuracy, proxies for response latency and word finding difficulties (namely, predictive distribution entropy or spread), confusion matrixes, and item specific effects (i.e., fitting a Bayesian model of what lexical features contributed to a correct or incorrect spoken word classification). We also analysed what form the networks' internal speech representations took, using a statistical physics method known as mean-field theory based manifold analysis to measure the average degree of spread of a single neural representation, and its overlap with competitor representations. These techniques are described in more detail below.

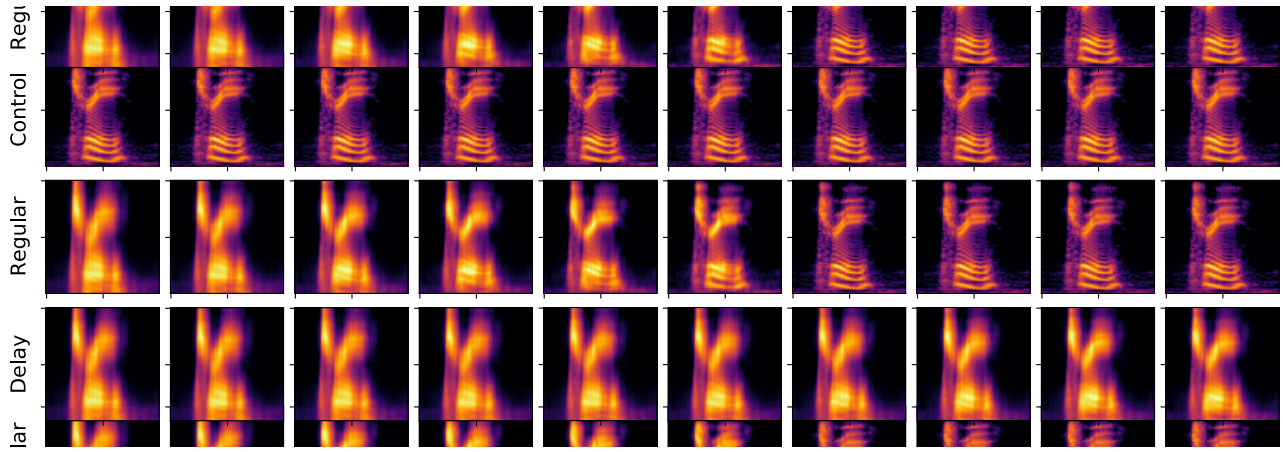
Convolutional neural networks are, in the vast majority of research, configured 'a-developmentally'. That is, parameters such as the number of layers or number of neurons per layer, etc. are fixed at the outset, and remain static during network training and testing (cf. Westermann et al., 2006; Westermann & Ruh, 2012; Westermann & Jones, 2021; these studies similarly involve neural networks that change structurally during learning, e.g., in

terms of the number of hidden units that they have). In contrast, one innovation of the current study was to model the maturation of high-resolution frequency discrimination skills using what is known as scheduled learning. That is, we ran distinct populations of neural networks in which frequency discrimination ability matured at different rates, according to different schedules across ten epochs of training. As can be seen in Figure 2, raw spoken word waveforms were initially passed through a cochleagram model developed specifically to replicate typical, human cochlea function (McDermott & Simoncelli, 2011). The resultant 100 × 100-dimension cochleagram images were then passed through the deep convolutional neural network and later into a 35-way classifier. In three discrete conditions we manipulated the maturation of that initial cochleagram model in three neural networks (n = 3, N = 9). Networks one, two, and three in each condition had identical weight initialisations. This ensured that any group-level performance discrepancies observed were not a function of the networks' starting states. Condition one was a-developmental – i.e., a baseline or control network – meaning that this network received high-resolution speech input from the outset and no changes to the network occurred during ten epochs of training (see Figure 3, row one). In contrast, the cochlea models of networks in conditions two and three maturated according to a specific schedule. In condition two, frequency resolution started low, but improved rapidly, resulting in full-resolution processing (i.e., baseline-equivalent acuity) by epoch seven (Figure 3, row two). This can be seen in the increasing y-axis acuity (i.e., decreasing vertical blur) across the cochleagrams in row two of Figure 3. Networks in condition three, in contrast, started with precisely the same standard of frequency resolution as the networks in condition two – that is, frequency resolution is identical during training epoch one in the regular and delay conditions – but then followed a delayed maturation schedule, never reaching baseline acuity (Figure 3, row 1). In both the delay and regular conditions, frequency resolution was constrained using a normalised box filter with a kernel of shape

$(1, y)$, where y decreased at different rates over ten epochs: from 25 to 1 in the regular condition and from 25 to 16 in the delay condition.

Figure 3

Schedules of Simulated Basilar Membrane Maturation



Note. Shown is a cochleagram of the word *tree* under varying rates of rates of maturation in spectral (i.e., *y*-axis) acuity within three conditions (control, regular, delay), and across ten cycles (epochs) of training.

Methods of analysis

All post-simulation analyses were conducted in R (RStudio Team, 2016). During training and testing, networks were presented with cochleagrams and in response output probability distributions over their 35-word lexicons. The word assigned the highest probability was taken as a network's classification and where this corresponded to the true target cochleagram a 'hit' was scored. The analysis of our training data involved measuring spoken word classification accuracy by training epoch. At test, we measured classification accuracy and the average maximum probability and probability distribution entropy output when a classification was made. These metrics provide a proxy for a network's certainty in its classifications. A high probability, low entropy (i.e., low spread) distribution signals high certainty in a judgement, while a low probability, high entropy (i.e., high spread) distribution signals low certainty in a judgement.

We then teased apart item-specific effects, looking for subsets of words on which regular or delayed networks performed better or worse. As part of this analysis into item-specific effects, we ran a Bayesian regression model (Burkner, 2017) in which the percentage of correct classifications per word was predicted by condition (i.e., regular, delayed) in interaction with two relevant independent variables that have generated considerable interest in developmental psycholinguistics: word frequency and word phonological neighbourhood density (e.g., Ambridge et al., 2015; Jones & Brandt, 2019; Rispens et al., 2015). Word frequency quantifies how common the word is in the exposure language, here the speech commands corpus from which training words were randomly sampled. Phonological neighbourhood density meanwhile quantifies the average distance, calculated on the basis of phonological transcriptions, between each word and the other 34 words in the training data. Relatively high input frequency is regularly associated with better language learning in children (Ambridge et al., 2015), while high phonological distance (i.e., phonemic dissimilarity) may improve speech classification accuracy among human listeners because potential between-item confusion is lower (Karimi & Diaz, 2020). As our modelling approach did not involve semantic representations it was not possible to include other variables of potential interest such as word concreteness, valence, or relevance to infants and babies (Braginsky et al., 2018; Jones & Brandt, 2019).

Artificial neural networks are sometimes criticised for being inscrutable ‘black boxes’. Yet, there exist numerous methods that enable the researcher to go beyond performance metrics such as accuracy alone to peer inside the network and understand how it is representing information in the service of completing a certain task. Exploiting such methods is vital to the current study because our interest is in how a processing hierarchy modelling the auditory pathway from brainstem to cortex optimises in the face of low-level constraints on frequency discrimination. Convolutional neural network activation patterns

have been shown to align broadly (i.e., not on a layer-to-structure level of granularity) with activation patterns in the biological brain (Kell et al., 2018; cf. Thompson, 2020). Furthermore, Bishop and MacArthur's work in this direction shows that even when there is apparently no group difference in performance metrics such as accuracy, frequency resolution deficits may be associated with different neural signatures across groups with and without language disorder (Bishop & McArthur, 2005; McArthur & Bishop, 2004). Similarly, Chonchaiya et al. (2013) showed that auditory brainstem responses continuous with immature brainstem optimisation predict relatively poor language outcomes. We wondered whether a similar neural signature of auditory processing impairments within the context of language learning deficits would emerge within our computational framework.

To better understand how our neural networks dynamically optimised to cochlea representations with varying spectral acuity (Figure 3), we used a recently developed framework known as mean field theory based manifold analysis (MFTMA; Figure 4; Chung et al., 2018; Chung & Abbott, 2021; Cohen et al., 2020). Under this approach, each neuron in any given structure of the auditory pathway, for instance the inferior colliculus, is configured as a single axis against which the spiking activity in that neuron can be plotted. Collectively, neurons in a given neural structure then define a neural state space (Figure 4A; graphically, a collection of axes) in which patterns of activation can be plotted either as trajectories through time or averaged spikes-per-second vectors. Given neural noise and variability in speaker and communicative context, no two instances of any given speech string stimulate the same response vector within that neural state space, i.e., repeated spoken instances of a given linguistic structure never stimulate each neuron in the state space to the same degree. Repeated exposure to a range of exemplars from a single linguistic class, whether phoneme, word, or construction, therefore stimulates a unified population response known as a 'manifold', which is a quasi-continuous subspace of the neural state space that can be

494 considered the neural basis of the representation of that class (Cohen et al., 2020). Implicitly
495 estimating the bounds of this neural manifold is considered integral to recognising and
496 producing novel yet valid speech, as if recognising that instances of this class may regularly
497 stimulate activation patterns within but not substantially outside this region of the state space
498 (Cohen et al., 2020; DiCarlo & Cox, 2007; Stephenson et al., 2020; Yamins & DiCarlo,
499 2016).

500 **Figure 4**

501 *Principles of Neural Population Geometry*

502

503 *Note.* Panel A shows the spoken words *tree*, *three*, and *two* as response vectors in high
504 dimensional space, with axes N_1 to N_n representing the response of a specific neuron within
505 the population in spikes per second. The population here could be any structure within the
506 auditory pathway (e.g., inferior colliculus, medial geniculate nucleus, etc.). Note that
507 response vectors can also be shown as trajectories over time (e.g., see Chung & Abbott,
508 2021). Exemplars of the same word, e.g., *tree*, reside in a different neural response vector as a
509 function of neural noise and speaker and context effects, but collectively form a quasi-
510 continuous manifold. (NB. In a deviation from the mathematical definition of a manifold,
511 neural manifolds need not be smooth and continuous, but are instead held to comprise the
512 convex hull of the distribution of neural responses elicited by a fixed class of stimulus.)
513 Panels B to D illustrate the neural basis of the well-studied transformation across the auditory
514 system from noise sensitive to speech selective responses (e.g., Davis & Johnsrude, 2003;
515 DeWitt & Rauschecker, 2012; Kaas et al., 1999; Okada et al., 2010). Early in the auditory

pathway manifolds of different speech strings intersect substantially due to cellular responsiveness to low-level auditory features. Intersecting manifolds are then incrementally untangled and reduced in dimensionality across the auditory pathway. Panel C shows an intermediate, 'low-capacity' system in which residual manifold tangling is evident. Panel D shows an optimal system with distributed speech representations that accommodate variability in the speech stream, but which are discrete and amenable to forming the focus of attention. The dotted line in panels C and D illustrates a simulated attentional mechanism (implicated in both recognition and retrieval) which is overwhelmed (Panel C) or effective (Panel D) as a function of the precision of activated long-term memories. Adapted from Jones and Westermann (2022).

The major contribution of the mean field theory based manifold analysis method is to enable us to treat distributed biological and artificial neural activation patterns as continuous geometric shapes that we can measure. Essentially, the convex hull of the collected response vectors (i.e., the points in Figure 4A) elicited by a fixed class of stimuli is treated as a single geometric object. In the current study, we are interested in two geometric quantities of neural representation that have received significant attention in the computational neuroscience literature. First, we are interested in the *dimensionality* of the pattern of activation (i.e., the manifold) underpinning responses to a certain class of spoken words (i.e., all instances of *tree*). That is, we are interested in how spread out through the neural state space speech representations are. Second, and relatedly, we are interested in the overlap between competitor neural representations, such as those underpinning the phonologically similar words *tree* and *three*. Within the MFTMA literature overlap is quantified in terms of *classification capacity*, which is derived by calculating the number of speech manifolds that can be linearly separated from all competitor representations and standardising the result by network layer size. In a low-capacity system representations are highly overlapping (i.e.,

discrete representations involve activity in shared neurons), and the system struggles to use a linear separator to recognise or retrieve any single representation given this overlap (Figure 4B, Figure 4C). In a high-capacity system, representation dimensionality (and other highly correlated quantities including manifold radius) has been reduced to a level at which overlap is low and linear separation is more straightforward (Figure 4D).

With these properties in mind, Jones and Westermann (2022) drew a parallel between variance in a network's classification capacity and the demands placed on human working memory or attentional systems as a function of the precision of activated long-term memories. Activated low-precision long-term memories, i.e., memories with high dimensionality, place high demands on the system and compromise efficient processing, overwhelming working memory and attention (Figure 4C). On the other hand, activated high-precision long-term memories, i.e., memories with low dimensionality, place low demands on the processing system, because procedures including speech recognition and retrieval are facilitated if the target representation is relatively discrete (Figure 4D).

Research in this area, both computational work and work involving humans, points to potentially domain general transformations in representational structure from low-level structures such as the auditory nerve to high-level structures such as the peripheral auditory cortex. Broadly, low-level structures are *noise sensitive*, and so manifolds show extensive overlap (i.e., high dimensionality representations in a low-capacity system). However, within both biological and artificial neural processing hierarchies, architectural features such as pooling functions (where, for instance, a neuron fires if *any* antecedent neuron fires) mean that early noise sensitive representations become increasingly *speech selective* (Davis & Johnsrude, 2003; DeWitt & Rauschecker, 2012; Kaas et al., 1999; Okada et al., 2010; Yamins & DiCarlo, 2016). That is, we go from high-dimension representations in a low-capacity system early in the pathway, to low-dimension representations in a high-capacity system late

in the pathway. The neural population geometry view of this trajectory is illustrated in Figure 4, panels B, C, and D. Jones and Westermann did not present a maturational account of frequency resolution and speech deficits. Instead, their interest was on explaining variance in working memory task performance. However, these authors did show that the trajectory shown in Figure 4 could be disrupted by the addition of broad Gaussian noise to input representations. Here we intend to build substantially on this work by (i) using cochleagrams developed expressly to simulate human auditory physiology, and (ii) manipulating cochleagrams during training in line with known trajectories in the maturation of frequency discrimination skills, something we believe to be unique to the current study.

It is worth noting that we are using a powerful neural network with a large number of training samples of a relatively small number of word types. In general, these are perfect conditions for training a highly robust neural network that copes well in the face of input noise. Our intention throughout this project was to keep our manipulation subtle in line with the notion of a possibly subtle derailment of a typical maturational trajectory. Indeed, looking at Figure 3 it is clear that the cochleagrams in epoch 10 retain something of a recognisable contour across conditions, and it might not be too challenging to visually identify this particular word, *tree*, from the cochleagrams of certain other words within the 35-word cohort. We did not, therefore, expect dramatic differential effects in the region of, for instance, 25% performance accuracy, which is the sort of disparity sometimes seen in empirical studies using so-called ‘extreme-group designs’, which compare quite severely language-impaired children to children with strong language skills (see West et al., 2017, for a criticism of this approach). Instead, we were looking for potentially subtle but consistent disparities in network optimisation indices and behaviour across conditions that align well with current behavioural and neurophysiological evidence from children with and without language learning difficulties.

This study was not preregistered. All data and materials required to re-run the simulations and analyses presented in this manuscript are available from the following repository: https://osf.io/x2h8k/.

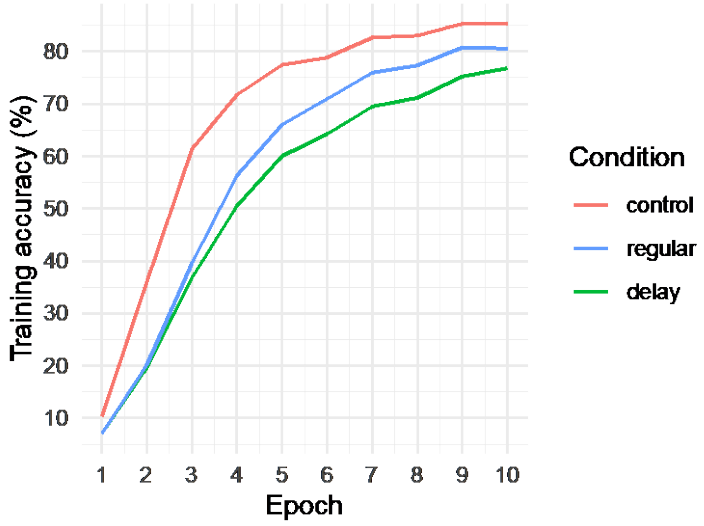
Results

Classification accuracy, probability, and entropy

In the analyses that follow, network performance is collapsed and reported as a condition mean. Spoken word classification accuracy by condition and training epoch is shown in Figure 5. Across epochs, networks in the optimal, a-developmental control condition outperformed the developmental networks in both regular and delay conditions. Constraining the maturation of high-resolution frequency discrimination according to the schedules shown in Figure 3 promoted a clear disparity between regular and delay networks, with the regular networks performing better after epoch two and this gap widening in line with the disparity in the resolution of spectral information generated by the networks' cochlea model (Figure 3).

Figure 5

Training Accuracy by Epoch and Condition



By epoch ten accuracy averaged 85.3% in the control condition, 80.6% in the regular condition, and 76.8% in the delay condition. A similar pattern was observed at test, where speech classification accuracy averaged 85.1% in the control condition, 83.9% in the regular condition, and 79.6% in the delay condition. During training and at test, accuracy reflects the networks' ability to correctly classify spoken word cochleagrams. The difference between these analyses is that training-phase accuracy describes a learning trajectory, while test-phase accuracy reflects a cross-sectional analysis that is conducted when training is complete.

The accuracy data above represents a record of hits as a proportion of total exposures. However, it is also possible to get a picture of the networks' confidence in their predictions by analysing the maximum probability assigned to a prediction and the entropy (or spread, in bits) of the probability distribution output. This analysis indicated greater uncertainty in the predictions made by networks in the developmental conditions than in the optimal condition, and greatest uncertainty in networks in the delay condition. Mean maximum probability assignment stood at 86.7% in the control condition, 81.5% in the regular condition, and 78.6% in the delay condition, while entropy or distribution spread in bits stood at 0.443 control, 0.612 regular, and 0.693 delay (i.e., indicating increasingly spread-out predictive distributions). A similar pattern was observed when limiting our analysis to hits only: Mean maximum probability assignment = 91.4% control, 87.2% regular, and 85.3% delay; entropy in bits = 0.306 control, 0.449 regular, and 0.496 delay.

In summary, networks in the maturational delay condition not only performed significantly less accurately than comparison networks, but also output relatively broad, highly spread probability distributions over their lexicons, considering many competitor words and assigning the true target relatively low probability even when accurate. Therefore, neural networks with maturational deficits in frequency resolution take longer to encode speech information, and metrics of test performance (i.e., low max probability, high entropy)

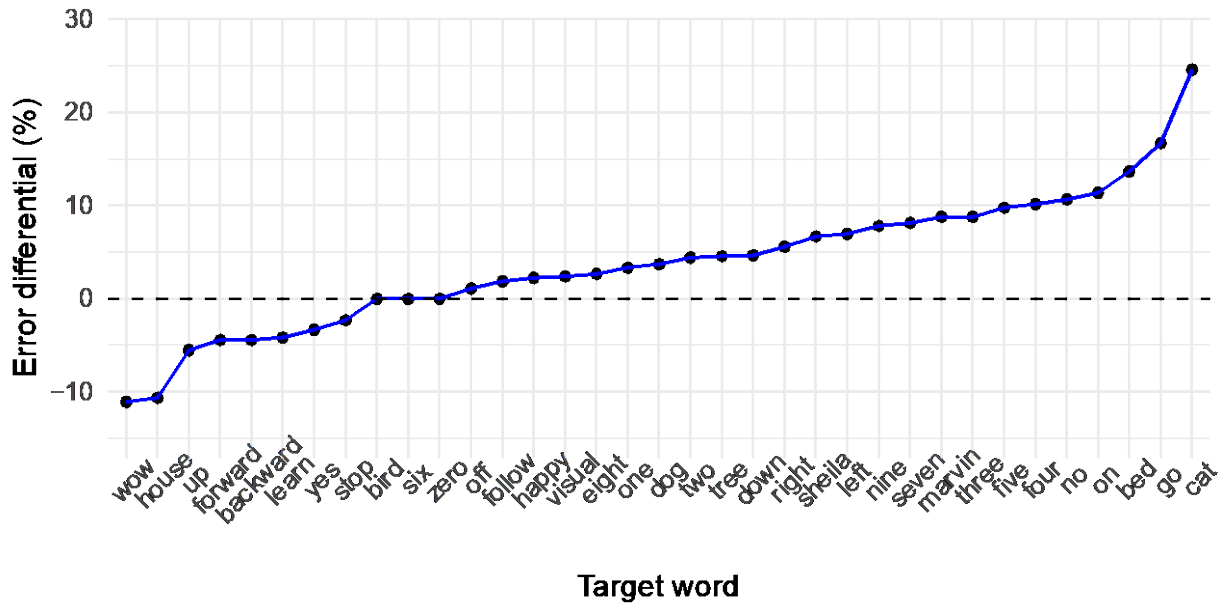
suggest that formed speech encodings are inefficiently organised. In response to speech input, more of what we might consider the networks' long-term memory (i.e., the fixed 35-word lexicon) becomes activated (i.e., we see high-spread predictive distributions), and the true target may be swamped in activated competitor representations. Qualitative analogies might be seen here between network performance and the DLD literature showing: (i) delayed acquisition profiles (Norbury et al., 2016; a parallel with the disparity in network accuracy over training epochs); (ii) lower spoken word recognition accuracy (Andreu et al., 2012; Evans et al., 2018; Rispens et al., 2015; Velez & Schwartz, 2010; a parallel with the network test-phase accuracy disparity), and; (iii) word finding difficulties and residual uncertainty even when performing accurately, as evidenced, for instance, in eye tracking paradigms (Kambanaros et al., 2015; McMurray et al., 2019; Messer & Dockrell, 2006; a parallel with high entropy, low probability activation patterns). Later, we examine the representational basis of these performance profiles. First, however, we aimed to determine the particular words that networks in the regular and delay conditions found difficult to encode and classify, as well as to understand why networks found these words difficult.

Item-specific effects

We began our item-specific analyses by computing a by-item accuracy differential, calculated by subtracting the average percentage accurate at test for each word in the delay condition from the average percentage accurate for each word in the regular condition. The result is shown in Figure 6. Here, a positive value indicates a performance advantage, as a percentage, for the regular network, and a negative value indicates a performance advantage for the delay network. Zero differential indicates no performance difference between conditions with respect to a particular word.

Figure 6

Item Accuracy Differential



Note. All 35 words from the speech commands dataset are shown along the *x*-axis. The error differential is shown on the *y*-axis. A positive differential value signals an advantage (as a percentage accurate) for the networks in the regular maturation condition. A negative differential value signals an advantage for the networks in the delay condition.

Networks in the regular condition outperformed networks in the delay condition with respect to 24 out of 35 words, sometimes reaching a differential of 24.6% (for the word *cat*). Networks in the delay condition, in contrast, performed better on eight words, with a maximum differential of -11.11% for the word *wow*. Clearly, then, error rates vary as a function of the target word. To better understand these effects, we looked at confusion matrices for predictions made during speech classification in each condition. The top ten most confused words in the regular and delay conditions are presented in Table 1 and Table 2 respectively. These tables show the true word, the total number of misclassifications of that word, the most common misclassification of that word, the number of times that the most common misclassification occurred, and most common misclassification as a proportion of total misclassifications (%).

Table 1

Top Ten Speech Classification Errors in the Regular Condition.

Word	Total misclassifications	Most common misclassification	Number	Proportion of total misclassifications (%)
tree	66	three	17	25.76
no	141	go	26	18.44
follow	54	four	7	12.96
go	78	no	10	12.82
up	72	off	9	12.5
house	75	off	8	10.67
four	69	forward	7	10.14
five	123	on	10	8.13
one	90	nine	7	7.78
off	93	on	7	7.53

Table 2

Top Ten Speech Classification Errors in the Delay Condition.

Word	Total misclassifications	Most common misclassification	Number	Proportion of total misclassifications (%)
tree	66	three	17	25.76
no	141	go	30	21.28
go	78	no	13	16.67
four	69	forward	10	14.49
five	123	on	16	13.01

on	132	five	16	12.12
right	108	five	10	9.26
two	114	go	10	8.77
three	114	eight	9	7.89
no	141	down	9	6.38

In many cases, the phonological overlap likely responsible for the misclassification is clear, for instance with respect to *tree* and *three* or *no* and *go*, and it is noteworthy that networks struggled by some margin with respect to these particular competitor words. Similar patterns are discussed by Karimi and Diaz (2020), who review classification disadvantages for near neighbours under certain experimental conditions. At first glance, then, networks appear to be broadly sensitive to similar spectral features input as human listeners (e.g., struggling with items like *tree* and *three*). Yet, Table 1 and Table 2 also illustrate examples which apparently deviate from this pattern, for instance the apparently high rates of misclassification of the word *five* as the word *on*, or the misclassification of the word *house* as *off*. It is difficult to imagine this pattern performance in human participants, and this may attest to the fact that despite the many gross similarities between processing in artificial neural networks and the human brain, artificial neural networks may attend to different features of the input in the service of reducing error in a given task. We return to this question below.

To further understand the above disparities in item accuracy between conditions we fitted a Bayesian regression model in which test phase accuracy (as a percentage) was predicted by standardised frequency and phonological distance, both in interaction with condition (i.e., regular, delay). We centred on frequency and phonological distance as predictor variables given their importance in the child language literature. However, alternative predictor variables of interest (e.g., orthographic word length) can be experimented with using the Jupyter notebook and R script associated with this project.

Frequency quantified the number of times that a word appeared in the randomly sampled training data. Meanwhile, phonological distance was computed as the mean optimal string alignment (OSA) distance between a phonological transcription of each target word and of all other words in the speech commands corpus.

A range of diagnostics showed that this simple regression model with a skew normal likelihood and weakly informative priors fitted well (i.e., rhats at 1.0, a large number of effective samples, and credible posterior predictive checks; see supplementary materials and the brms documentation for further details; Burkner, 2017). Figure 7 shows the estimates from our Bayesian model.

Figure 7

Estimates from a Bayesian Model of the Influence of Frequency and Phonological Similarity on Speech Classification Accuracy

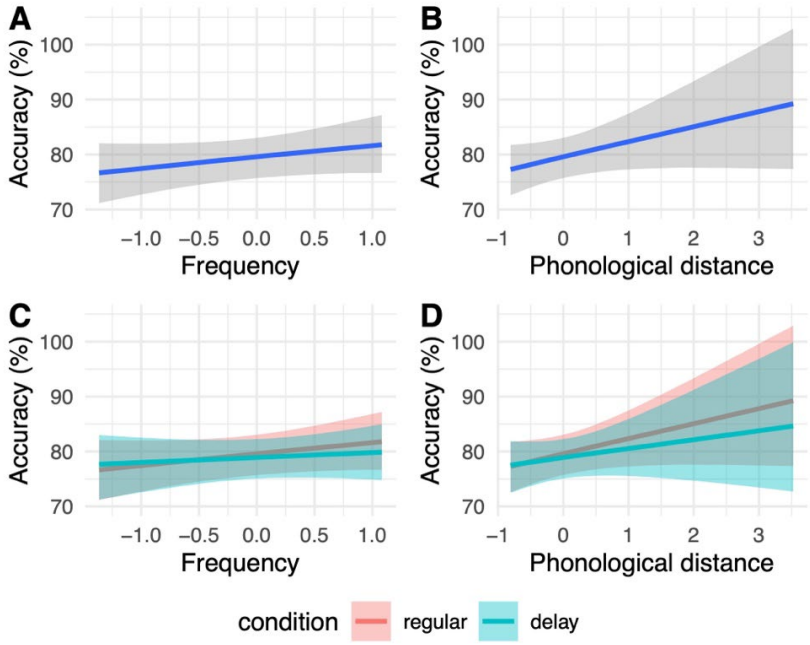


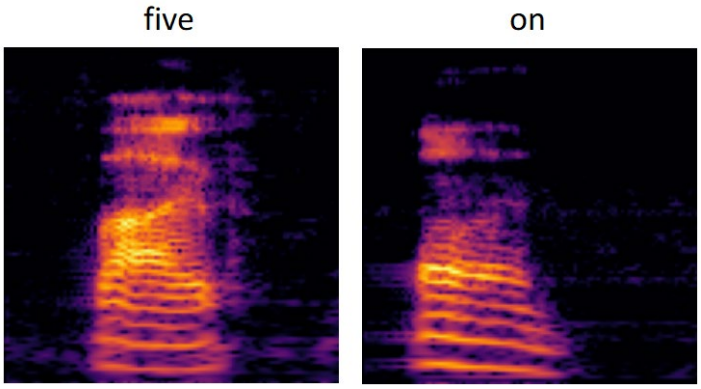
Figure 7, panels A and B show that across groups, classification accuracy was on average higher for high frequency (β = 2.11; 95% CI = -0.97 to 5.45), phonologically distinctive (β = 2.82; 95% CI = -0.46 to 6.46) words. While the credible intervals (CIs) associated with these estimates cross zero, indicating that zero may be the true effect, a substantial proportion of probability mass is positively assigned, suggesting that a positive association is likely. Meanwhile, Figure 7, panels C and D show that these effects interact slightly with condition but tend in the same positive direction (see R code for full estimates: https://osf.io/x2h8k/). In each case, networks with rapidly maturing high-resolution cochlea models benefitted slightly more from high frequency and greater phonologically distinctiveness.

In summary, item specific analyses indicate that while networks struggled to different degrees with different words, they nevertheless struggled with broadly similar features of the dataset, misclassifying close competitor words such as *tree* and *three* most frequently and performing best when words were highly frequent in the training data and phonologically

distinctive. Higher resolution low-level auditory representations enabled networks in the regular condition to better exploit these input statistics. These results may be expected given that at any particular period the regular and delay networks sit at different points on the same developmental trajectory. The resulting performance profiles are in agreement with the general observation that the language of children with DLD is delayed rather than deviant (Kan & Windsor, 2010; see also *Discussion*). That is, the language of children with DLD is often similar to that of younger children with typical language skills (though see Bishop, 2014a). That said, our item-specific analysis also revealed potential discrepancies between artificial neural network and human performance. For instance, we observed a high rate of misclassification of exemplars of *five* and *on* (see also the *house* and *off* misclassification rate), which at face value would appear unlikely in human participants. If, however, we look at representative raw cochleagrams of the words *five* and *on*, for instance, these classification errors perhaps make more sense (all cochleagrams can be visualised using the associated scripts). The distributions of energy in the exemplars shown in Figure 8 are at least visually quite similar, and would of course be even more similar were we depreciate their acuity across the *y*-axis (for reference, compare the spectral profiles of *five* and *on* to the quite different profile shown for *tree* in Figure 3).

Figure 8

Reprsentative Cochleagrams of the Words 'Five' and 'On'



Viewing Figure 8, it may appear reasonable that an artificial neural network would misclassify degraded instances of *five* and *on*. But how about a human? Of potential relevance when considering this question is a large research literature looking at so-called adversarial examples. These are stimuli which, when noise that is typically imperceptible to humans is added, result in the radical misclassification of those stimuli in an otherwise high-performing network (Goodfellow et al., 2014; of course, the *y*-axis blur in our study is perceptible to humans). For instance, an image of a panda with visually imperceptible noise added to it may be misclassified as a gibbon. Understanding adversarial examples is a vital part of research on human and machine learning alignment, because it throws light on the marginal disparities between biological and artificial systems that in many other ways appear to perform similarly. Intriguingly, there is limited evidence that the same adversarial examples that derail artificial neural network classification may also affect human performance, just to a lesser extent and emerging in metrics of classification confidence such as response time rather than in raw error rates (Elsayed et al., 2018). Two possibilities, then, are that either the *five* and *on* misclassification error and similar striking errors seen in the current simulations are evidence the inescapable disparity between artificial and biological auditory perceptual processing systems, or, on the other hand, that we might be able to elicit similar patterns of classification behaviour (e.g., extended response times) in humans using similar stimuli. There is a precedent for this type of work in the domain of visual processing (Elsayed et al., 2018) but a similar experiment in the domain of auditory processing was outside the scope of the current project.

Visualising internal representations – Mean field theory based manifold analyses

The cochlea models that provide input to the deep convolutional neural networks used in these simulations were scheduled to mature according to one of two developmental time courses. In contrast, the neural networks into which cochleagrams were passed were provided

765 with a randomised initial weight matrix, which was matched across networks and conditions,
766 but which then optimised freely to solve the specific problems of speech encoding,
767 recognition, and retrieval. (Note that the control network presents an optimal system which is
768 free to optimise in the absence of any significant low-level constraint.) The performance
769 profiles detailed above – specifically the disparities in accuracy, probability, entropy, and
770 item specific effects – point to systematic differences in dynamic optimization that, given
771 matching across networks, can result only from these low-level maturational constraints in
772 high-resolution frequency processing. We are, therefore, modelling discrepancies in optimal
773 adaptation in the face of different low-level constraints. But what does optimisation in the
774 face of a low-level frequency discrimination deficit look like? To better understand the
775 optimisation profiles of networks in our three conditions, and therefore to unpick the
776 representational basis of the performance discrepancies seen in networks across these
777 conditions, we turned to mean field theory based manifold analyses.

778 Variables of primary interest were (i) manifold dimensionality and (ii) classification
779 capacity. Manifold dimensionality quantifies how spread out through a neural state space
780 long-term speech representations are – i.e., how many artificial neurons (as a proportion of
781 the layer size) are implicated in the representation of that speech string. Classification
782 capacity quantifies the number of speech manifolds that can be linearly separated from all
783 competitor representations, again standardised by network layer size. Analysis of biological
784 and artificial neural networks suggests that dimensionality decreases across the auditory and
785 visual perceptual systems, and accordingly, that system capacity increases (Chung et al.,
786 2018; Chung & Abbott, 2021; DiCarlo & Cox, 2007). This transformation reflects the gradual
787 de-noising of neural representations in a perceptual hierarchy. Speech representations, for
788 instance, are shown to become decreasingly noise sensitive and increasingly speech selective
789 during transformation from the basilar membrane to the peripheral auditory cortex and

beyond (Davis & Johnsrude, 2003; DeWitt & Rauschecker, 2012; Kaas et al., 1999; Okada et al., 2010).

System classification capacity has been interpreted as a measure of not only representation overlap, but also of attention or working memory load, given that calculating classification capacity involves linearly discriminating discrete representations from the system's 'long-term memory' in a manner continuous with cognitive recognition and retrieval (Jones & Westermann, 2022). This view is in line with so-called state based frameworks in which working memory is understood as activated long-term memory that must be optimised to 'fit' within an attentional spotlight (Adams et al., 2018; Oberauer, 2013, 2019). Importantly, reducing manifold dimensionality in order to boost system classification capacity is a product of training in a given task, here speech encoding and classification. Training with the same data in a different task, for instance speaker recognition, would result in an internal network structure optimised for this task (i.e., activation patterns forming manifolds of speaker voice characteristics; Stephenson et al., 2020). The result of this task-specific optimization process is presented in Figure 9, which shows the average manifold dimensionality and classification capacity in networks' penultimate layers antecedent to the classifier (see Figure 2) as a function of training epoch.

Figure 9

Changes in Manifold Dimensionality and Classification Capacity During Training

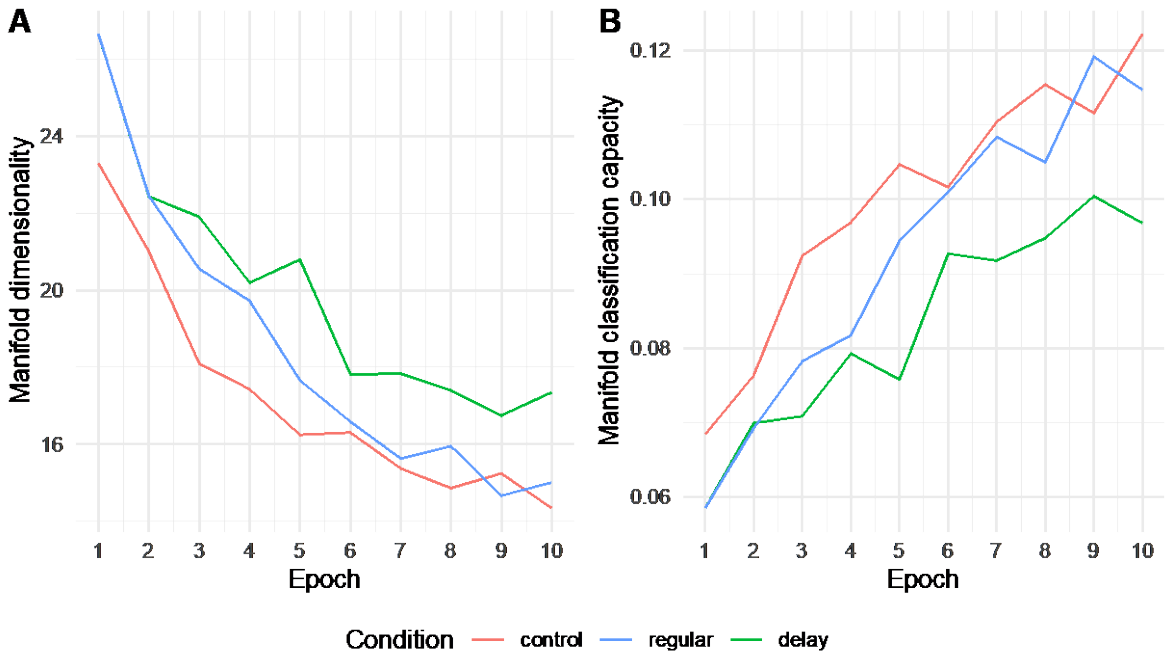
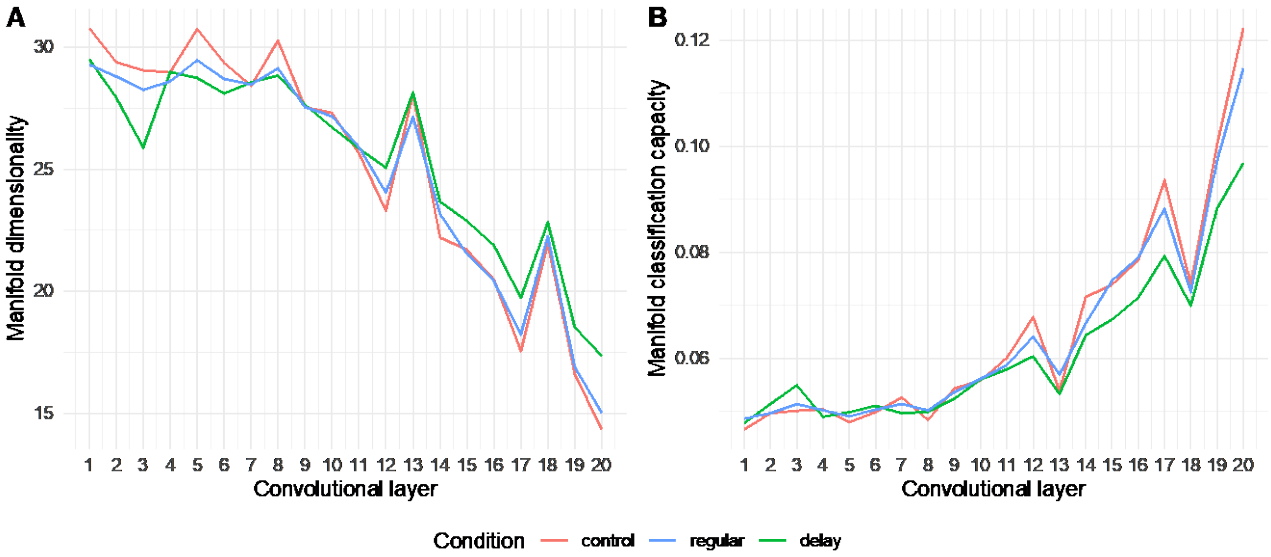


Figure 9 shows a clear disparity in the optimisation of internal speech representations across conditions. Over ten epochs, networks following the regular cochlea maturation schedule increasingly approached control standards of optimisation supporting low-dimensional representation (Figure 9A). In contrast, despite an overall decrease across epochs, the average dimensionality of internal spoken word representations formed in networks in the delay condition remained significantly higher, i.e., these representations were substantially more 'spread out' in a relatively poorly optimised neural state space (Figure 9A). Figure 9, panel B shows that this inability to optimize efficiently and reduce manifold dimensionality had a severe effect on the delay networks' ability to retrieve any single representation from their internal 'long-term memory' systems – what we interpret here as a form of simulated working memory or attentional capacity deficit. In essence, the delay networks optimised to noise, and this means that the artificial neural response patterns underpinning the long-term representations of different spoken words intersect substantially, making efficient recognition and retrieval difficult. Graphically, it is as though the delay networks remain in the suboptimal state shown in Figure 4C, rather than approaching the relatively optimal state shown in Figure 4D alongside networks in the regular and control conditions.

The same representational disparity can be seen post-training across the networks' layers. In Figure 10 we show the previously reported trajectory (e.g., Yamins & DiCarlo, 2016) across the auditory processing hierarchy from high-dimensional manifolds in a low-capacity system to low-dimensional manifolds in a high-capacity system. Again, this reflects the system optimising to render initially noise sensitive representations (i.e., waveform representations containing speaker effects, etc.) increasingly speech selective (i.e., word type representations in the 35-word lexicon). There is, however, a clear optimisation disparity between networks in the delay condition and networks in the control and regular conditions in terms of both dimensionality and classification capacity at higher levels of the processing hierarchy. This again demonstrates that due to maturational constraints in the cochlea model, networks in the delay condition failed to learn those spectral features of the speech input that are essential to effective speech encoding, recognition, and retrieval, with noise permeating the system and attentional capacity overwhelmed accordingly (Figure 10B). This can be seen most clearly in Figure 10 with respect to layers 19 and 20, where delay networks deviate sharply from the regular and control networks with respect to both dimensionality and classification capacity.

Figure 10

Manifold Dimensionality and Classification Capacity Across the Layers of Trained Networks

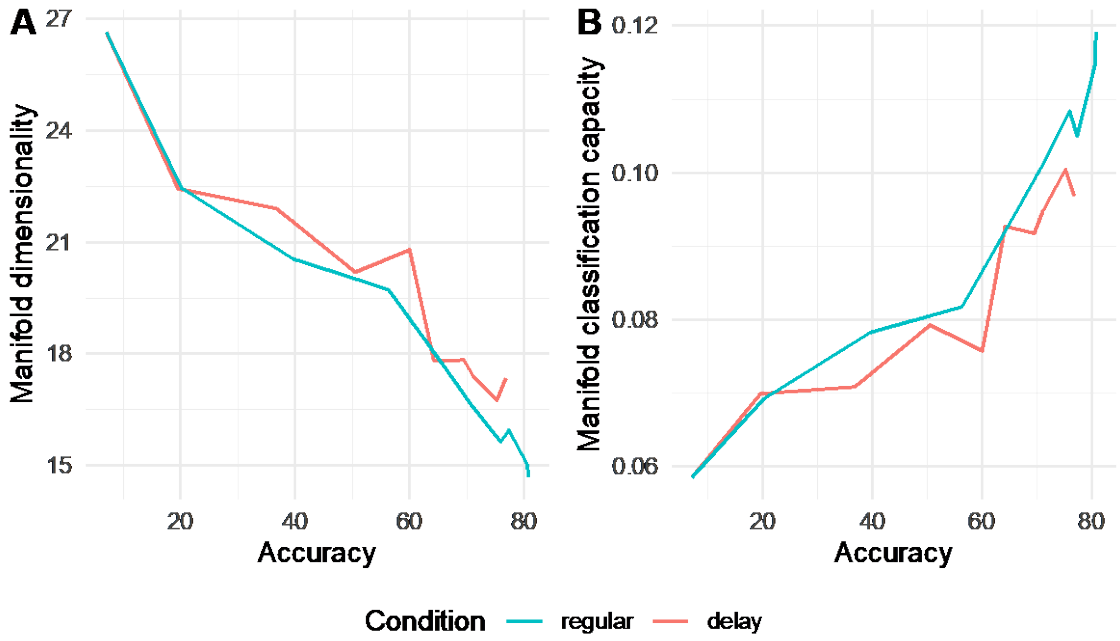


Finally, we observed optimisation disparities even when regular and delay networks were matched on performance accuracy. In Figure 11 we show manifold dimensionality and manifold classification capacity as a function of training-phase accuracy, by network condition. As in Figure 9, manifold dimensionality and classification capacity are computed for the networks' final convolutional layer (see Appendix), which is antecedent to the 35-way classifier. Despite very occasional overlap, manifold dimensionality is high and classification capacity is low in the delayed networks relative to the regular networks even when networks in these conditions perform with similar accuracy. This result demonstrates the importance of scrutinising the internal representations that artificial neural networks form. Based on accuracy alone we may have wrongly inferred that networks were achieving that level of performance in the same task in the same way, overlooking important differences in the standards of internal optimisation. The finding of representational deficits despite matched levels of performance echoes Bishop and McArthur's reports of electrophysiological discrepancies between children with and without DLD even when DLD-group performance is at threshold (Bishop & McArthur, 2005; McArthur & Bishop, 2004; see also Mengler et al., 2005) and Chonchaiya et al.'s (2013) evidence that signatures of poor auditory brainstem optimisation are predictive of language outcomes. This reaffirms the important point that

apparent successes in task performance may not be underpinned by similar qualities of learning, a point also made by McMurray et al. (2012).

Figure 11

Manifold Dimensionality and Classification Capacity by Performance Accuracy



In summary, these simulations illustrate how dynamic adaptation to biologically plausible models of cochlea function that mature at different rates results in different optimization profiles, which underpin disparities in key performance metrics (i.e., accuracy, max probability assignment, and entropy) and which are evident despite performance accuracy matching (Figure 11). By constraining the development of high-resolution frequency discrimination, we curtailed the systems' ability to optimise to encode the key spectral features of the speech input that are integral to solving the task at hand, namely speech recognition and retrieval. The performance of networks in the delayed condition in this study makes the prediction that the optimization profile of a biological speech encoding system with a low-level frequency discrimination deficit will show high dimensional speech representations (i.e., relatively dispersed neural activation patterns on exposure to speech stimuli) which intersect with competitor speech representations, and which are, therefore, not

amenable to forming an effective focus of attention. Apparent attention deficits then emerge as a result of being thinly spread rather than atypically capacity limited. Prior work involving typically developing adults has shown that this prediction regarding divergent neural activation patterns is in principle testable in language disordered populations (Davis & Johnsrude, 2003; DeWitt & Rauschecker, 2012; Kaas et al., 1999; Okada et al., 2010). Indeed, as described in our literature review, there is already some evidence from language disordered populations that is broadly continuous with this claim. For instance, low quality, 'fuzzy', speech representations are well documented in the behavioural literature looking at children with DLD (Claessen et al., 2009, 2013; Claessen & Leitão, 2012a, 2012b), and atypical neurophysiological signatures indicating suboptimal auditory pathway optimisation that is predictive of language impairment have been reported in a number of studies (Bishop & McArthur, 2005; Chonchaiya et al., 2013; McArthur & Bishop, 2004).

Discussion

Frequency discrimination deficits are widely recognised among children with language learning difficulties (Bishop & McArthur, 2005; McArthur & Bishop, 2004; Mengler et al., 2005). Yet, the nature of these deficits and their relation to speech processing problems remain unclear. The neural microarchitecture supporting high resolution frequency discrimination matures from the prenatal period through to later childhood, and it is possible that the frequency discrimination deficits seen among some children with language learning difficulties stems from a disruption to this typical developmental trajectory (Bishop & McArthur, 2005; McArthur & Bishop, 2004). Given that frequency tuning throughout the auditory pathway is predominantly attributable to the structural properties of the basilar membrane (i.e., the membrane's *mechanical* gradient, including fiber diameter, density, and regularity; Tani et al., 2021), we hypothesised that the protracted maturation of the structural properties of the basilar membrane may provide a good starting point for inquiry into the

source of frequency discrimination deficits in children with neurodevelopmental disorder. Disruption to the structure of the basilar membrane has been demonstrated empirically in animal models manipulating emilin 2 expression, which results in a deficient mechanical gradient and therefore suboptimal functioning of the auditory pathway not supporting high-resolution frequency processing (Amma et al., 2003; Russell et al., 2020).

We developed this theoretical account through a series of computational simulations of speech encoding, recognition, and retrieval. The networks used in these simulations incorporated inner ear models developed to replicate human cochlea function (McDermott & Simoncelli, 2011) that were fed into deep convolutional neural networks. Despite many important differences, for instance in scale, complexity, and the use of undifferentiated cell types, deep convolutional neural networks have demonstrated significant correspondences with human behavioural and neural responses across a range of tests of audition including speech localization, pitch perception, and hearing in noise (Francl & McDermott, 2022; Kell et al., 2018; Saddler et al., 2021). Our own innovation was to configure the cochlea models that formed a fundamental component of our networks to mature according to different developmental trajectories (i.e., baseline or optimal, regular, and delayed), and to analyse how the subsequent auditory-linguistic pathway optimised in the service of speech encoding, recognition, and retrieval.

Our analysis of networks in the delayed cochlea maturation condition qualitatively replicated the linguistic behaviour and neurophysiology of individuals with language learning difficulties in a number of ways, showing: (i) delayed acquisition profiles (Norbury et al., 2016); (ii) lower spoken word recognition accuracy (Andreu et al., 2012; Evans et al., 2018; Rispens et al., 2015; Velez & Schwartz, 2010); (iii) word finding and retrieval difficulties and uncertainty even when performing accurately, as evidenced, for instance, in eye tracking paradigms (i.e., Kambanaros et al., 2015; McMurray et al., 2019; Messer & Dockrell, 2006);

(iv) ‘fuzzy’ long-term speech representations (Claessen et al., 2009, 2013; Claessen & Leitão, 2012a, 2012b) and neurophysiological signatures of immature neural optimisation that are associated with speech and language difficulties (Bishop & McArthur, 2005; Chonchaiya et al., 2013; McArthur & Bishop, 2004); and (v) apparent working memory and attention deficits that are attributable, we believe, to the imprecision of long-term speech representations (Gray et al., 2019; Henry & Botting, 2017; Jones & Westermann, 2022). Our results illustrate that optimising to low-level, low-resolution spectral representations significantly curtails the capacity of the system to form speech representations supporting efficient recognition and retrieval.

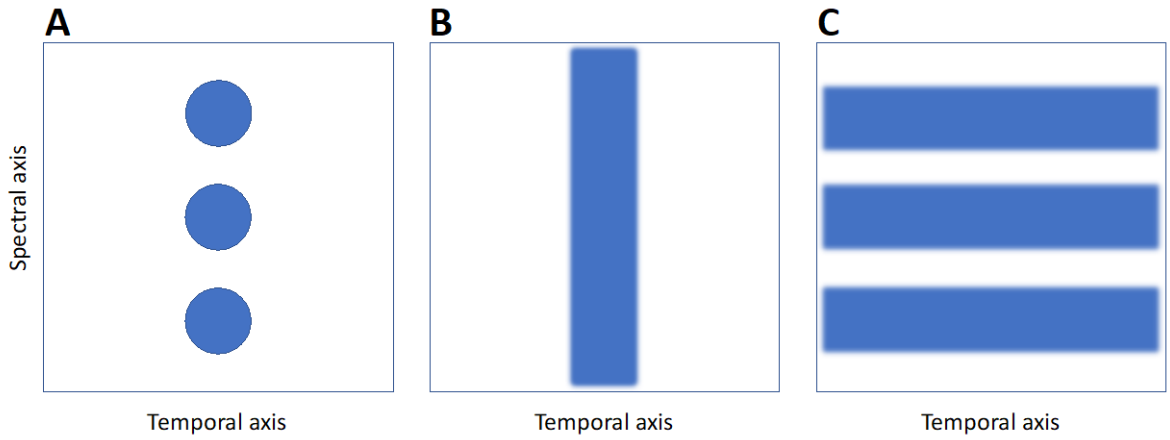
We see, then, that some of the mechanisms widely thought to play a causal role in speech and language disorder may ‘come for free’ if we assume a low-level frequency discrimination deficit. This includes not only the hypothesised working memory capacity bottleneck (Archibald & Gathercole, 2006), which dominates DLD research but which we have argued to be a possible epiphenomenon (see also Jones & Westermann, 2022), but also the so-called lateral inhibition deficit suggested by McMurray et al. (2019). McMurray et al. (2019) argue that a key feature of early language disorder may be an inability to inhibit activated competitor representations during speech recognition in retrieval. Our simulations suggest, however, that an apparent lateral inhibition deficit may be an emergent characteristic of a suboptimal auditory processing hierarchy. Networks in the delayed cochlea maturation condition of our simulations uniformly output predictive distributions with high spread (i.e., high entropy) and low maximum probability assignment, signalling heightened uncertainty and broader activation of the lexicon in response to speech stimuli. As in the case of the hypothesised working memory capacity limitation, then, we believe that evidence offered in support of a deficit in a functionally discrete lateral inhibition mechanism may instead reflect

target isolation being overwhelmed due to the imprecision of activated long-term speech representations; a process illustrated in Figure 4C.

It may be argued that the results presented in the current study were inevitable. That is, that disrupting the quality of the cochlea representations that networks could form would necessarily lead to worse performance. But this is not the case. Indeed, data disruption, for instance blurring, skewing, re-colouring, or clipping the training data is regularly used in machine learning, where the process is termed 'data augmentation', to boost network performance by preventing overfitting and attenuating attention to consistent features (Chollet, 2021). The discrepancies in network performance seen in the current study are, therefore, attributable to the specific features that we degraded – i.e., frequency information distributed across the *y*-axis – being essential to the efficient encoding and therefore recognition and retrieval of natural speech. Feature importance is graphically illustrated in Figure 12. In Panel A we show three dots, exemplifying schematic features that may help us to classify a particular stimulus. In our case the dots in Figure 12 represent the distinctive frequency components of a speech string. If, as seen in Panel B, we were to degrade this stimulus across the *y*-axis (i.e., the frequency dimension) this would – as demonstrated in the current study – cause problems in determining the identity of that stimulus. On the other hand, degrading the same stimulus across the *x*-axis (i.e., the temporal dimension) preserves the stimulus' critical features.

Figure 12

Degrading Critical Features



That is not to say that the *x*-axis degradation seen in Figure 12 Panel C, has no effect. Indeed, work by Saddler et al. (2021) and Saddler and McDermott (2022) has shown that manipulating auditory nerve firing rates to degrade temporal information has a significant negative effect on sound localisation and voice recognition. The point is, then, that when it comes to encoding speech efficiently specifically for the purposes of accurate recognition and retrieval, low-level auditory representations with high-resolution, discrete frequency components appear essential. And, as we have highlighted throughout this article, there is good evidence that high-resolution frequency discrimination is a core problem among some children with language learning difficulties.

The above discussion of the concept of feature importance may bring some light to the debate regarding whether the auditory processing deficits seen among some children with neurodevelopmental disorders are spectral (i.e., frequency-based) or temporal in nature. As discussed in our introduction, the early dominant view in DLD research was that the performance deficits seen are temporal in nature, but this view has weakened considerably in the face of failed replications (Strong et al., 2011; Bishop & McArthur, 2005; McArthur & Bishop, 2004; see Rosen, 2003, for review). In contrast, there is compelling evidence that the auditory processing deficits seen among some children with language problems are spectral in nature (Bishop & McArthur, 2005; McArthur & Bishop, 2004; Mengler et al., 2005). Computational simulation indicates that both spectral and temporal information are crucial to

effective speech processing, but that the relative importance of these cues is differentially weighted as a function of the task. Temporal acuity is vital, for instance, in the context of voice recognition and sound localisation (Saddler et al., 2021; Saddler & McDermott, 2022). Yet when it comes to encoding speech for the purposes of recognition and retrieval, the current simulations show that high frequency component acuity is key.

It may also be argued that, had we allowed the cochlea models of our delayed networks to continue maturing until they reach the same standard as the cochlea models of our regular networks, network optimisation and therefore task performance may have eventually normalised. This is true, and reflects the fact that artificial neural networks are not bound by any hard and fast sensitive period or maturational constraints on physiology[3]. Language problems are, in contrast, often evident across the lifespan, suggesting long-lasting disparities in the organisation of neural substrates supporting audition and speech. If we take a maturational view of frequency discrimination and speech and language deficits, then, the critical questions are when and how the typical dynamic adaptation of the auditory pathway becomes 'frozen' in a sub-optimal state. This appears particularly puzzling given that the auditory pathway is, in general, highly plastic, for instance often adapting quickly to the fitting of a cochlear implant (e.g., Wang et al., 2021). One possibility is that the mechanical gradient of the basilar membrane (and, therefore, tonotopic sensitivity in membrane-posterior structures) never reaches optimal differentiation, as in our delayed networks. However, the locus of deficit may of course reside in any of the structures posterior to the cochlea that also support tonotopic mapping. For instance, Bishop and McArthur (2005) note that while the cochlea is typically fully developed by full-term birth, the auditory brainstem and subsequent structures continue to adapt through childhood, with frequency discrimination skills

[3] That said, we note that sensitive periods may stem from entrenchment rather than biological ossification, and can therefore emerge in computational systems (Thomas & Johnson, 2006).

improving accordingly. Bishop and McArthur (2005) hypothesise, therefore, that either (i) the delayed optimisation of higher-level structures within the auditory pathway, including the auditory cortex, may be protracted and then plateau with the onset of puberty, or (ii) that structures of the auditory pathway that support high-resolution frequency tuning may develop slowly but nevertheless fully, yet the cost of a protracted period of maturation during the initial phases of language development may be long lasting.

In this study, we have demonstrated how the auditory linguistic pathway may optimise in the face of a cochlea maturation deficit. The basilar membrane remains in our view a good starting point for future inquiry, because the deficits we see among children with DLD are spectral in nature and because the basilar membrane is the seat of tonotopic organisation throughout the auditory pathway. We also hypothesised that, given that emilin 2 plays a key role in the emergence of the development of the mechanical gradient of the basilar membrane (Amma et al., 2003; Russell et al., 2020; Tani et al., 2021), potential disruption to the expression of this gene might be considered (though we cite the emilin 2 literature primarily to emphasise how a genetic abnormality can in principle disrupt the emergence of the mechanical gradient of the basilar membrane). Yet, given the enormous complexity of the auditory pathway, numerous possibilities obviously remain. If, through empirical testing, a maturational account is ruled out, it will be necessary to look beyond an early 'freezing' of typical cochlea, auditory brainstem, and auditory cortex maturation, and to instead identify deviances in auditory pathway develop that could give rise to low-resolution frequency processing, for instance testing for mid-frequency sensorineural hearing loss (i.e., 'cookie-bite' hearing loss; see Ahmadmehrabi et al., 2022, for an adult study) that signals problems with the cochlea or auditory nerve, or identifying cortical dysplasia in neural substrates supporting audition and speech (Bishop, 2014b).

An important feature of the current study was to let our networks develop over time, using cochlea models that output representations of increasing spectral acuity according to different maturational trajectories (Figure 3). This developmental approach to modelling with neural networks is uncommon, though it is continuous with a limited number of connectionist studies that have let their networks develop as a function of experience (e.g., Elman, 1993; Westermann et al., 2006; Westermann & Ruh, 2012). We believe that such an approach is integral to the study of the developing brain and mind. Similar work is being conducted by Skelton (2022), who has developed a filter to simulate changes in the visual system during the neonatal period and infancy, which can be used in both experimental stimulus design and in computational models of neuro-cognitive development. This development-driven approach to computational modelling is likely to provide us with a much richer understanding of the emergence of human cognitive behaviour, relative to methods fundamentally aligned with a-developmental adult norms.

Like any method the use of artificial neural networks to understand human brain function and behaviour has its limitations. Neural networks are, of course, a dramatic simplification of the structure of the human brain, involving drastically fewer cells of identical, undifferentiated types, with activation functions allowing the communication of real numbers. What is more, biological and artificial neural networks learn differently. For instance, biological neural networks appear not to need thousands of labelled exemplars in order to learn spoken words (Lake et al., 2013; though see Lillicrap et al., 2020, for how the brain might approximate the backpropagation algorithm used in our neural networks). These architectural and algorithmic differences may underpin different performance profiles – the high misclassification rates with respect to *five* and *on* in our data might be a case in point here. Nevertheless, gross parallels between human performance and brain function and deep neural network activation patterns and performance have been observed repeatedly (Kell et

al., 2018; McDermott & Simoncelli, 2011; Saddler et al., 2021; Yamins & DiCarlo, 2016), and a reasonable qualitative mapping with the empirical data in the current study further supports this approach.

Modelling of the form presented here of course constitutes a counterpart to, and not replacement of, human assessment. Modelling forces us to be explicit about our assumptions, and – as we have demonstrated – may provide computational insight into the nature of representation, recognition, and retrieval within dynamic systems that have optimised to different fundamental constraints. Of course, further analysis involving humans is vital. There have already been important steps in this direction, with Chonchaiya et al. (2013) showing that neural signatures of immature auditory brainstem organisation are indicative of poorer language outcomes – a finding highly in agreement with the hypothesis developed in the current paper. To date, however, many studies of children with a diagnosis of DLD have included only rudimentary auditory assessments involving, for instance, backward masking, mismatch negativity, or glide discrimination, which can show significant variability before around eight years of age (Bishop et al., 2005; Bishop & McArthur, 2005; Sutcliffe et al., 2006). One particularly elegant example of the inadequacy of such approaches comes from research demonstrating that children diagnosed with attention deficit hyperactivity disorder (ADHD) can complete pure tone discrimination tasks when taking their medication but not when off their medication (Sutcliffe et al., 2006). This highlights the susceptibility of such tasks to non-auditory perceptual influences, including attention. Given the ubiquity of apparent auditory processing problems not only among children diagnosed with DLD but also across other early neurodevelopmental disorders such as developmental dyslexia, there is strong justification for a large-sample study involving rich early auditory assessments (including, for instance, extended high-frequency audiometry), longitudinal neuroimaging, and the assessment of later language outcomes.

The speech commands dataset was chosen for this project because it is free and openly available, and because it is unique in comprising such a large number of natural speech exemplars. One limitation of this resource, however, is that it comprises only 35 word types, meaning that only limited insight can be drawn from our item-specific analyses. While we believe that the use of the speech commands dataset in the current project is well justified, going forward it would be useful to replicate our findings using a larger dataset. In particular, it would be valuable to test children and artificial neural networks using the same speech stimuli, which could be recorded specifically for this purpose. This would support a relatively direct comparison between child and artificial neural network behaviour. Indeed, using this approach it would be possible to simulate real-world language interventions and to determine the computational basis of their efficacy.

Conclusion

Frequency discrimination is a core problem for many children with language learning difficulties, and through computational simulation we have shown how this deficit would propagate problems with the encoding, recognition, and retrieval of natural speech. Our simulations provide proof of concept that the optimisation of the auditory-linguistic pathway to low-resolution cochlea representations – part of a typical maturational trajectory that may be protracted in DLD – result in patterns of linguistic behaviour that align qualitatively with a range of empirical findings observed among children with DLD. Our speculation that the locus of such deficits may be a disruption to the maturation of the basilar membrane during a sensitive period of auditory pathway optimisation reflects the fact that the mechanical gradient of the basilar membrane provides the basis for the emergence of frequency sensitivity across the auditory-linguistic pathway. Yet, this hypothesis of course requires empirical testing. The auditory-linguistic pathway is a highly complex system which could be disrupted at any level. Also in need of further scrutiny is our speculation, given the

contemporary animal model literature, that atypicalities in emelin 2 expression may be implicated in the disruption of the emergence of the mechanical gradient of the basilar membrane (i.e., the development of fibril microarchitecture supporting high resolution processing, which promulgates the required tonotopic sensitivity through the auditory nerve, brainstem, and cortex). We fully recognise these elements of our argument to be speculation, albeit empirically driven speculation. Our view is simply that the weight of empirical evidence with respect to structural changes in the basilar membrane suggests that this hypothesis constitutes a strong starting point for further inquiry into the nature of auditory processing deficits in children with language learning difficulties.

References

Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of working memory: Differences in definition, degree of modularity, role of attention, and purpose. *Language, Speech, and Hearing Services in Schools*, *49*(3), Article 3. https://doi.org/10.1044/2018_LSHSS-17-0114

Ahmadmehrabi, S., Li, B., Epstein, D. J., Ruckenstein, M. J., & Brant, J. A. (2022). How Does the "Cookie-Bite" Audiogram Shape Perform in Discriminating Genetic Hearing Loss in Adults? *Otolaryngology–Head and Neck Surgery*, *166*(3), 537–539. https://doi.org/10.1177/01945998211015181

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. In *Journal of Child Language*. https://doi.org/10.1017/S030500091400049X

Amma, L. L., Goodyear, R., Faris, J. S., Jones, I., Ng, L., Richardson, G., & Forrest, D. (2003). An emilin family extracellular matrix protein identified in the cochlear basilar membrane. *Molecular and Cellular Neuroscience*, *23*(3), 460–472. https://doi.org/10.1016/S1044-7431(03)00075-7

Andreu, L., Sanz-Torrent, M., & Guàrdia-Olmos, J. (2012). Auditory word recognition of nouns and verbs in children with specific language impairment (SLI). *Journal of Communication Disorders*, *45*(1), Article 1. https://doi.org/10.1016/j.jcomdis.2011.09.003

Archibald, L. M. D., & Gathercole, S. E. (2006). Short-term and working memory in specific language impairment. *International Journal of Language and Communication Disorders*, *41*(6), Article 6. https://doi.org/10.1080/13682820500442602

Archibald, L. M. D., & Harder Griebeling, K. (2016). Rethinking the connection between working memory and language impairment. *International Journal of Language & Communication Disorders*, *51*(3), Article 3. https://doi.org/10.1111/1460-6984.12202

Astle, D. E., Holmes, J., Kievit, R., & Gathercole, S. E. (2022). Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry*, *63*(4), Article 4. https://doi.org/10.1111/jcpp.13481

Barman, A., Prabhu, P., Mekhala, V. G., Vijayan, K., & Narayanan, S. (2021). Electrophysiological findings in specific language impairment: A scoping review. *Hearing, Balance and Communication*, *19*(1), 26–30. https://doi.org/10.1080/21695717.2020.1807277

Bishop, D. V. M. (2006). What causes specific language impairment in children? *Current Directions in Psychological Science*, *15*(5), Article 5. https://doi.org/10.1111/j.1467-8721.2006.00439

Bishop, D. V. M. (2014a). Problems with tense marking in children with specific language impairment: Not how but when. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), Article 1634. https://doi.org/10.1098/rstb.2012.0401

Bishop, D. V. M. (2014b). *Uncommon Understanding (Classic Edition)*. Psychology Press. https://doi.org/10.4324/9780203381472

Bishop, D. V. M., Adams, C. V., Nation, K., & Rosen, S. (2005). Perception of transient nonspeech stimuli is normal in specific language impairment: Evidence from glide discrimination. *Applied Psycholinguistics*, *26*, 175–194. https://doi.org/10.1017.S0142716405050137

Bishop, D. V. M., Bishop, S. J., Bright, P., James, C., Delaney, T., & Tallal, P. (1999). Different origin of auditory and phonological processing problems in children with

language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(1), Article 1. https://doi.org/10.1044/jslhr.4201.155

Bishop, D. V. M., Hardiman, M. J., & Barry, J. G. (2012). Auditory Deficit as a Consequence Rather than Endophenotype of Specific Language Impairment: Electrophysiological Evidence. *PLoS ONE*, *7*(5), Article 5. https://doi.org/10.1371/journal.pone.0035851

Bishop, D. V. M., & McArthur, G. M. (2005). Individual differences in auditory processing in specific language impairment: A follow-up study using event-related potentials and behavioural thresholds. *Cortex*, *41*(3), Article 3. https://doi.org/10.1016/S0010-9452(08)70270-3

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2016). CATALISE: A multinational and multidisciplinary delphi consensus study. Identifying language impairments in children. *PLOS ONE*, *11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2018). *Consistency and variability in word learning across languages*. https://doi.org/10.31234/osf.io/cg6ah

Burkner, P.-C. (2017). *brms: Bayesian Regression Models using 'Stan'*. 154.

Burnham, D., & Mattock, K. (2014). Auditory development. In G. Bremner & T. Wachs (Eds.), *The Wiley Blackwell Handbook of Infant Development* (2nd ed., pp. 83–121).

Chollet, F. (2021). *Deep Learning with Python, Second Edition*.

Chonchaiya, W., Tardif, T., Mai, X., Xu, L., Li, M., Kaciroti, N., Kileny, P. R., Shao, J., & Lozoff, B. (2013). Developmental trends in auditory processing can provide early predictions of language acquisition in young infants. *Developmental Science*, *16*(2), 159–172. https://doi.org/10.1111/desc.12012

1197 Chung, S., & Abbott, L. F. (2021). Neural population geometry: An approach for
1198 understanding biological and artificial neural networks. *Current Opinion in*
1199 *Neurobiology*, *70*, 137–144. https://doi.org/10.1016/j.conb.2021.10.010
1200 Chung, S., Lee, D. D., & Sompolinsky, H. (2018). Classification and Geometry of General
1201 Perceptual Manifolds. *Physical Review X*, *8*(3), Article 3.
1202 https://doi.org/10.1103/PhysRevX.8.031003
1203 Claessen, M., Heath, S., Fletcher, J., Hogben, J., & Leitão, S. (2009). Quality of phonological
1204 representations: A window into the lexicon? *International Journal of Language and*
1205 *Communication Disorders*, *44*(2), Article 2.
1206 https://doi.org/10.1080/13682820801966317
1207 Claessen, M., & Leitão, S. (2012a). Phonological representations in children with SLI. *Child*
1208 *Language Teaching and Therapy*, *28*(2), Article 2.
1209 https://doi.org/10.1177/0265659012436851
1210 Claessen, M., & Leitão, S. (2012b). The relationship between stored phonological
1211 representations and speech output. *International Journal of Speech-Language*
1212 *Pathology*, *14*(3), Article 3. https://doi.org/10.3109/17549507.2012.679312
1213 Claessen, M., Leitão, S., Kane, R., & Williams, C. (2013). Phonological processing skills in
1214 specific language impairment. *International Journal of Speech-Language Pathology*,
1215 *15*(5), Article 5. https://doi.org/10.3109/17549507.2012.753110
1216 Cohen, U., Chung, S. Y., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry
1217 of object manifolds in deep neural networks. *Nature Communications*, *11*(1), Article
1218 1. https://doi.org/10.1038/s41467-020-14578-5
1219 Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language
1220 comprehension. *Journal of Neuroscience*, *23*(8), 3423–3431.
1221 https://doi.org/10.1523/jneurosci.23-08-03423.2003

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(8), 505–514. https://doi.org/10.1073/pnas.1113427109

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), Article 8. https://doi.org/10.1016/j.tics.2007.06.010

Echteler, S. M., Arjmand, E., & Dallos, P. (1989). Developmental alterations in the frequency map of the mammalian cochlea. *Nature*, *341*(6238), 147–149. https://doi.org/10.1038/341147a0

Elmahallawi, T. H., Gabr, T. A., Darwish, M. E., & Seleem, F. M. (2021). Children with developmental language disorder: A frequency following response in the noise study. *Brazilian Journal of Otorhinolaryngology*. https://doi.org/10.1016/j.bjorl.2021.01.008

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), Article 1.

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). *Adversarial Examples that Fool both Computer Vision and Time-Limited Humans*. https://doi.org/10.48550/ARXIV.1802.08195

Evans, J. L., Gillam, R. B., & Montgomery, J. W. (2018). Cognitive Predictors of Spoken Word Recognition in Children With and Without Developmental Language Disorders. *Journal of Speech, Language, and Hearing Research*, *61*(6), Article 6. https://doi.org/10.1044/2018_JSLHR-L-17-0150

Fletcher-Watson, S. (2022). Transdiagnostic research and the neurodiversity paradigm: Commentary on the transdiagnostic revolution in neurodevelopmental disorders by Astle et al. *Journal of Child Psychology and Psychiatry*, *63*(4), Article 4. https://doi.org/10.1111/jcpp.13589

Francl, A., & McDermott, J. H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, *6*(1), 111–133. https://doi.org/10.1038/s41562-021-01244-z

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and Harnessing Adversarial Examples*. https://doi.org/10.48550/ARXIV.1412.6572

Gray, S., Fox, A. B., Green, S., Alt, M., Hogan, T. P., Petscher, Y., & Cowan, N. (2019). Working memory profiles of children with dyslexia, developmental language disorder, or both. *Journal of Speech, Language, and Hearing Research*, *62*(6), Article 6. https://doi.org/10.1044/2019_JSLHR-L-18-0148

Haake, C., Kob, M., Willmes, K., & Domahs, F. (2013). Word stress processing in specific language impairment: Auditory or representational deficits? *Clinical Linguistics and Phonetics*, *27*(8), Article 8. https://doi.org/10.3109/02699206.2013.798034

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition*.

Henry, L. A., & Botting, N. (2017). Working memory and developmental language impairments. *Child Language Teaching and Therapy*, *33*(1), Article 1. https://doi.org/10.1177/0265659016655378

Hestvik, A., Epstein, B., Schwartz, R. G., & Shafer, V. L. (2022). Developmental Language Disorder as Syntactic Prediction Impairment. *Frontiers in Communication*, *6*, 637585. https://doi.org/10.3389/fcomm.2021.637585

Higgins, I., Stringer, S., & Schnupp, J. (2017). Unsupervised learning of temporal features for word categorization in a spiking neural network model of the auditory brain. *PLOS ONE*, *12*(8), e0180174. https://doi.org/10.1371/journal.pone.0180174

Jensen, J. K., & Neff, D. L. (1993). Development of Basic Auditory Discrimination in Preschool Children. *Psychological Science*, *4*(2), 104–107. https://doi.org/10.1111/j.1467-9280.1993.tb00469.x

Jones, S. D., & Brandt, S. (2019). Do children really acquire dense neighbourhoods? *Journal of Child Language*, *46*(6), 1260–1273. https://doi.org/10.1017/S0305000919000473

Jones, S. D., & Westermann, G. (2022). Under-resourced or overloaded? Rethinking working memory and sentence comprehension deficits in developmental language disorder. *Psychological Review*, *Advance online publication.* http://dx.doi.org/10.1037/rev0000338

Kaas, J. H., Hackett, T. A., & Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology*, *9*(2), 164–170. https://doi.org/10.1016/S0959-4388(99)80022-1

Kambanaros, M., Michaelides, M., & Grohmann, K. K. (2015). Measuring word retrieval deficits in a multilingual child with SLI: Is there a better language? *Journal of Neurolinguistics*, *34*, 112–130. https://doi.org/10.1016/j.jneuroling.2014.09.006

Kan, P. F., & Windsor, J. (2010). Word Learning in Children With Primary Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *53*(3), Article 3. https://doi.org/10.1044/1092-4388(2009/08-0248)

Karimi, H., & Diaz, M. (2020). When phonological neighborhood density both facilitates and impedes: Age of acquisition and name agreement interact with phonological neighborhood during word production. *Memory & Cognition*, *48*(6), 1061–1072. https://doi.org/10.3758/s13421-020-01042-4

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), Article 3. https://doi.org/10.1016/j.neuron.2018.03.044

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C.

(2016). Jupyter Notebooks—A publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016*, 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Lake, B. M., Salakhutdinov, R. R., & Tenenbaum, J. (2013). One-shot learning by inverting a compositional causal process. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 26). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/52292e0c763fd027c6eba6b8f494d2eb-Paper.pdf

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346. https://doi.org/10.1038/s41583-020-0277-3

Lopez-Poveda, E. A. (2014). Development of Fundamental Aspects of Human Auditory Perception. In *Development of Auditory and Vestibular Systems* (pp. 287–314). Elsevier. https://doi.org/10.1016/B978-0-12-408088-1.00010-5

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc.

McArthur, G. M., & Bishop, D. V. M. (2004). Which People with Specific Language Impairment have Auditory Processing Deficits? *Cognitive Neuropsychology*, *21*(1), 79–94. https://doi.org/10.1080/02643290342000087

McArthur, G. M., & Bishop, D. V. M. (2005). Speech and non-speech processing in people with specific language impairment: A behavioural and electrophysiological study. *Brain and Language*, *94*(3), Article 3. https://doi.org/10.1016/j.bandl.2005.01.002

McDermott, J. H., & Simoncelli, E. P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron*, *71*(5), 926–940. https://doi.org/10.1016/j.neuron.2011.06.032

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), Article 4. https://doi.org/10.1037/a0029872

McMurray, B., Klein-Packard, J., & Tomblin, J. B. (2019). A real-time mechanism underlying lexical deficits in developmental language disorder: Between-word inhibition. *Cognition*, *191*, 104000. https://doi.org/10.1016/j.cognition.2019.06.012

Mengler, E. D., Hogben, J. H., Michie, P., & Bishop, D. V. M. (2005). Poor frequency discrimination is related to oral language disorder in children: A psychoacoustic study. *Dyslexia*, *11*(3), 155–173. https://doi.org/10.1002/dys.302

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal Processing Deficits of Language-Learning Impaired Children Ameliorated by Training. *Science*, *271*(5245), Article 5245. https://doi.org/10.1126/science.271.5245.77

Messer, D., & Dockrell, J. E. (2006). Children's naming and word-finding difficulties: Descriptions and explanations. *Journal of Speech, Language, and Hearing Research*, *49*(2), Article 2. https://doi.org/10.1044/1092-4388(2006/025)

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry*, *57*(11), Article 11. https://doi.org/10.1111/jcpp.12573

Novitski, N., Huotilainen, M., Tervaniemi, M., Näätänen, R., & Fellman, V. (2007). Neonatal frequency discrimination in 250–4000-Hz range: Electrophysiological evidence.

Clinical Neurophysiology, *118*(2), 412–419. https://doi.org/10.1016/j.clinph.2006.10.008

Nuttall, A. L., Ricci, A. J., Burwood, G., Harte, J. M., Stenfelt, S., Cayé-Thomasen, P., Ren, T., Ramamoorthy, S., Zhang, Y., Wilson, T., Lunner, T., Moore, B. C. J., & Fridberger, A. (2018). A mechanoelectrical mechanism for detection of sound envelopes in the hearing organ. *Nature Communications*, *9*(1), 4175. https://doi.org/10.1038/s41467-018-06725-w

Oberauer, K. (2013). The focus of attention in working memory—From metaphors to mechanisms. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00673

Oberauer, K. (2019). Working memory and attention – A conceptual analysis and review. *Journal of Cognition*, *2*(1), Article 1. https://doi.org/10.5334/joc.58

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., Serences, J. T., & Hickok, G. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, *20*(10), 2486–2495. https://doi.org/10.1093/cercor/bhp318

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library*.

Pinker, S. (1994). *The language instinct* (1st ed). W. Morrow and Co.

Python Software Foundation. (2008). Python. In *Python Language Reference* (No. 3). https://www.python.org/

Rispens, J., Baker, A., & Duinmeijer, I. (2015). Word recognition and nonword repetition in children with language disorders: The effects of neighborhood density, lexical

frequency, and phonotactic probability. *Journal of Speech, Language, and Hearing Research*, *58*(1), Article 1. https://doi.org/10.1044/2014_JSLHR-L-12-0393

Rosen, S. (2003). Auditory processing in dyslexia and specific language impairment: Is there a deficit? What is its nature? Does it explain anything? *Journal of Phonetics*, *31*(3–4), Article 3–4. https://doi.org/10.1016/S0095-4470(03)00046-9

RStudio Team. (2016). RStudio: Integrated development for R. *[Online] RStudio, Inc., Boston, MA URL Http://Www. Rstudio. Com*. https://doi.org/10.1007/978-81-322-2340-5

Russell, I. J., Lukashkina, V. A., Levic, S., Cho, Y.-W., Lukashkin, A. N., Ng, L., & Forrest, D. (2020). Emilin 2 promotes the mechanical gradient of the cochlear basilar membrane and resolution of frequencies in sound. *Science Advances*, *6*(24), eaba2634. https://doi.org/10.1126/sciadv.aba2634

Saddler, M., Gonzalez, R., & McDermott, J. H. (2021). Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, *12*(1), 7278. https://doi.org/10.1038/s41467-021-27366-6

Saddler, M., & McDermott, J. (2022, March 19). *The role of temporal coding in everyday hearing: Evidence from deep neural networks* [Poster]. https://www.world-wide.org/cosyne-22/role-temporal-coding-everyday-hearing-ba39ae72/

Shafer, V. L., Morr, M. L., Kreuzer, J. A., & Kurtzberg, D. (2000). Maturation of Mismatch Negativity in School-Age Children: *Ear and Hearing*, *21*(3), 242–251. https://doi.org/10.1097/00003446-200006000-00008

Skelton, A. (2022, August 25). *A digital filter to simulate infant visual experience*. Lancaster Conference on Infant and Early Child Development (LCICD), Lancaster, UK.

Stephenson, C., Feather, J., Padhy, S., Elibol, O., Tang, H., McDermott, J., & Chung, S. (2020). *Untangling in Invariant Speech Recognition*. http://arxiv.org/abs/2003.01787

Strong, G. K., Torgerson, C. J., Torgerson, D., & Hulme, C. (2011). A systematic meta-analytic review of evidence for the effectiveness of the 'Fast ForWord' language intervention program. *Journal of Child Psychology and Psychiatry*, *52*(3), Article 3. https://doi.org/10.1111/j.1469-7610.2010.02329.x

Sumner, C. J., Wells, T. T., Bergevin, C., Sollini, J., Kreft, H. A., Palmer, A. R., Oxenham, A. J., & Shera, C. A. (2018). Mammalian behavior and physiology converge to confirm sharper cochlear tuning in humans. *Proceedings of the National Academy of Sciences*, *115*(44), 11322–11326. https://doi.org/10.1073/pnas.1810766115

Sutcliffe, P. A., Bishop, D. V. M., Houghton, S., & Taylor, M. (2006). Effect of Attentional State on Frequency Discrimination: A Comparison of Children With ADHD On and Off Medication. *Journal of Speech, Language, and Hearing Research*, *49*(5), 1072–1084. https://doi.org/10.1044/1092-4388(2006/076)

Tallal, P. (2013). Fast ForWord®. In *Progress in Brain Research* (Vol. 207, pp. 175–207). Elsevier. https://doi.org/10.1016/B978-0-444-63327-9.00006-0

Tallal, P., Stark, R., Kallman, C., & Mellits, D. (1981). A Reexamination of Some Nonverbal Perceptual Abilities of Language-Impaired and Normal Children as a Function of Age and Sensory Modality. *Journal of Speech, Language, and Hearing Research*, *24*(3), 351–357. https://doi.org/10.1044/jshr.2403.351

Tani, T., Koike-Tani, M., Tran, M. T., Shribak, M., & Levic, S. (2021). Postnatal structural development of mammalian Basilar Membrane provides anatomical basis for the maturation of tonotopic maps and frequency tuning. *Scientific Reports*, *11*(1), 7581. https://doi.org/10.1038/s41598-021-87150-w

Tharpe, A. M., & Ashmead, D. H. (2001). A Longitudinal Investigation of Infant Auditory Sensitivity. *American Journal of Audiology*, *10*(2), 104–112. https://doi.org/10.1044/1059-0889(2001/011)

Thomas, M. S. C., & Johnson, M. H. (2006). The computational modeling of sensitive periods. *Developmental Psychobiology*, *48*(4), 337–344. https://doi.org/10.1002/dev.20134

Thompson, J. A. F. (2020). *Characterizing and comparing acoustic representations in convolutional neural networks and the human auditory system* [PhD Thesis, Université de Montréal]. https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/24665/Thompson_Jessica_2020_these.pdf?sequence=2

Ullman, M. T., & Pierpont, E. I. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex*, *41*(3), Article 3. https://doi.org/10.1016/S0010-9452(08)70276-4

Velez, M., & Schwartz, R. G. (2010). Spoken word recognition in Sshool-age children with SLI: Semantic, phonological, and repetition priming. *Journal of Speech, Language, and Hearing Research*, *53*(6), Article 6. https://doi.org/10.1044/1092-4388(2010/09-0042)

Wang, S., Lin, M., Sun, L., Chen, X., Fu, X., Yan, L., Li, C., & Zhang, X. (2021). Neural Mechanisms of Hearing Recovery for Cochlear-Implanted Patients: An Electroencephalogram Follow-Up Study. *Frontiers in Neuroscience*, *14*, 624484. https://doi.org/10.3389/fnins.2020.624484

Warden, P. (2018). *Speech commands: A dataset for limited-vocabulary speech recognition*. http://arxiv.org/abs/1804.03209

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science*, *May 2016*, Article May 2016. https://doi.org/10.1111/desc.12552

Westermann, G., & Ruh, N. (2012). A neuroconstructivist model of past tense development and processing. *Psychological Review*, *119*(3), 649–667. https://doi.org/10.1037/a0028258

Westermann, G., Sirois, S., Shultz, T. R., & Mareschal, D. (2006). Modeling developmental cognitive neuroscience. *Trends in Cognitive Sciences*, *10*(5), 227–232. https://doi.org/10.1016/j.tics.2006.03.009

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), Article 3. https://doi.org/10.1038/nn.4244

Appendix

ResNet-18 convolutional layer specification and hyperparameters

Layer index	Layer name	Output size	Kernel size	Stride	Padding
1	Conv. 2D	1, 64	7, 7	2, 2	3, 3
2	Conv. 2D	64, 64	3, 3	1, 1	1, 1
3	Conv. 2D	64, 64	3, 3	1, 1	1, 1
4	Conv. 2D	64, 64	3, 3	1, 1	1, 1
5	Conv. 2D	64, 64	3, 3	1, 1	1, 1
6	Conv. 2D	64, 128	3, 3	2, 2	1, 1
7	Conv. 2D	128, 128	3, 3	1, 1	1, 1
8	Conv. 2D	64, 128	1, 1	2, 2	n/a
9	Conv. 2D	128, 128	3, 3	1, 1	1, 1
10	Conv. 2D	128, 128	3, 3	1, 1	1, 1
11	Conv. 2D	128, 256	3, 3	2, 2	1, 1
12	Conv. 2D	256, 256	3, 3	1, 1	1, 1
13	Conv. 2D	128, 256	1, 1	2, 2	n/a
14	Conv. 2D	256, 256	3, 3	1, 1	1, 1
15	Conv. 2D	256, 256	3, 3	1, 1	1, 1
16	Conv. 2D	256, 512	3, 3	2, 2	1, 1
17	Conv. 2D	512, 512	3, 3	1, 1	1, 1
18	Conv. 2D	256, 512	1, 1	2, 2	n/a
19	Conv. 2D	512, 512	3, 3	1, 1	1, 1
20	Conv. 2D	512, 512	3, 3	1, 1	1, 1

Note. See Jupyter notebook for full network specification.

Hyperparameters

Optimizer: stochastic gradient descent
Learning rate: .001
Momentum: .9
Loss function: cross-entropy loss

1456