Simulating Perceptual Tasks With Deep Neural Networks to Improve Diagnostics of Hearing Impairment

Jens C Thuren Lindahl, Bastian Epp, Hyojin Kim, Gerard Encina-Llamas

Abstract—Measuring pure-tone thresholds is the gold standard clinical tool for assessing the health of the auditory system. However, several studies on both animals and humans show impairment and hearing deficits that are not reflected by the absolute threshold, leading to the term ”hidden hearing loss (HHL)”. HHL has been shown in humans using electrophysiological measurements; however, these come at a high expense and are not applicable for clinical use. The need for a behavioural test is therefore of high interest. Such tests have been suggested based on a heuristic approach. However, this method might lead to a non-optimal trial-and-error strategy. Using a state-of-art model of the auditory nerve (AN) paired with a deep neural network (DNN) model, we investigate gap detection as a method for detecting cochlear synaptopathy (CS). Furthermore, we propose this approach as a general framework for investigating behavioural tests before conducting expensive and time consuming human experiments. We trained the DNN model on natural speech data and simulated a broadband-noise (BBN) gap detection task. The trained model was sensitive to CS and hearing threshold shift induced by inner hair cell (IHC) dysfunction. In contrast, the DNN model achieved lower gap detection thresholds (GDTs) with induced outer hair cell (OHC) dysfunction. Our results suggests gap detection as a behavioural test sensitive to CS and potentially also to IHC dysfunction.

1 Introduction

Pure-tone audiometry is the gold standard diagnostic test for the assessment of hearing in the clinics and has remained as such since the beginning of the previous century (Jones and Knudsen, 1925). Pure-tone audiometry is a behavioural test that estimates the absolute hearing threshold and it is used to diagnose the severity and type of hearing loss. This is the basis for aural rehabilitation, and thus the primary measure for hearing aid fitting. A significant elevation of hearing thresholds relative to the reference normal hearing (NH) hearing thresholds are usually associated with a damage in the periphery of the auditory system. However, since the mid-twentieth century, it is known that hearing thresholds do not represent all peripheral damage well. For instance, damage to AN neurons are not associated with significant threshold elevations assessed behaviourally in cats (Schuknecht and Woellner, 1955). More recently, selective damage to IHCs inflicted using ototoxic drugs in chinchillas showed that the hearing threshold was not significantly altered for IHC losses up to about 80 % (Lobarinas et al., 2013). In contrast, it is well known that damage to OHCs is very well associated with pure-tone threshold elevation (Ryan and Dallos, 1975).

It has been estimated that between 5 to 10 % of the patients in the clinics who show normal hearing thresholds, self-report hearing difficulties particularly in challenging acoustic environments (Saunders and Haggard, 1989; Kumar et al., 2007; Hind et al., 2011; Tremblay et al., 2015). These patients may suffer from hidden pathologies not revealed as a threshold elevation in the pure-tone audiogram, thus the termed ”HHL” was coined (Schaette and McAlpine, 2011). More precisely, a loss of cochlear synapses in the absence of hair cell damage, named CS, has been reported to be undetectable by pure-tone audiometry (Kujawa and Liberman, 2009).

To overcome such limitation of behavioural measures, the feasibility of electrophysiological measures for detecting sensory loss has been investigated. CS is one of the most well-studied types of HHL and has been demonstrated in several animal models, such as mice (Kujawa and Liberman, 2009; Furman et al., 2013; Shaheen et al., 2015; Parthasarathy and Kujawa, 2018), rats (Lobarinas et al., 2017), chinchillas (Hickox et al., 2017; Hickman et al., 2018), guinea pigs (Lin et al., 2011; Liu et al., 2012) and rhesus macaques (Valero et al., 2017). CS has also been demonstrated in humans (Makary et al., 2011; Viana et al., 2015; Wu et al., 2019, 2020, 2021). In non-human animal studies, it was shown that the loss of cochlear synapses did not alter hearing thresholds as assessed through auditory brainstem responses (ABR) and distortion product otoacoustic emission (DPOAE) thresholds. Nevertheless, synaptic losses resulted in a permanent reduction of supra-threshold responses in the ABR (Kujawa and Liberman, 2009) and the envelope following responses (EFR) (Shaheen et al., 2015; Parthasarathy and Kujawa, 2018). Several studies have investigated the use of similar auditory evoked responses in humans (e.g., Bharadwaj et al., 2015; Mehraei et al., 2016; Bramhall et al., 2017; Liberman et al., 2016; Prendergast et al., 2017; Guest et al., 2017; Fulbright et al., 2017; Encina-Llamas et al., 2019; Bramhall et al., 2021; Encina-Llamas et al., 2021; Maele et al., 2021; Märcher-Rørsted et al., 2022). Unfortunately, these studies led to inconclusive results (Bramhall et al., 2019), showing no clear effects in different groups of listeners suffering from different degrees of CS, presumably. Although CS has been associated with perceptual deficits, its impact on perception remains unclear too (Plack et al., 2014).

Compared to behavioural tests, electrophysiological techniques are more time-consuming, more expensive and require advanced equipment and specialised operators. These deficits limit their applicability in a clinical setting and motivate the need for a behavioural test that is sensitive to HHL. Previous studies have compared electrophysiological measures presumably sensitive to CS with some behavioural tasks, such as amplitude modulation (AM) detection (Bharadwaj et al., 2015), finding very weak correlations. The decision of choosing one task versus another was based on a heuristic interpretation of the function of the peripheral auditory system. For example, Bharadwaj et al. proposed an AM detection task with the hypothesis that higher degrees of CS would degrade the detection of shallower modulated tones. However, this hypothesis may not be consistent with how AM depth is encoded at the level of the AN, as suggested by physiological models of the auditory periphery (Encina-Llamas et al., 2019, 2021).

The evaluation of any newly proposed behavioural tests suggested after a heuristic interpretation of the function of the hearing system could lead to a non-optimal trial-and-error strategy. This evaluation would require testing on dozens of listeners, which is excessively time-consuming and expensive. An alternative could be a computational approach. The development of a framework where proposed behavioural tests could be simulated and evaluated regarding its sensitivity to CS is of high interest. Using such a framework would be a much more flexible and time-efficient approach, as different parameters could be adjusted in the theoretical computer framework without extensive cost, and the feasibility of the test could be evaluated prior to testing on actual participants.

Typically, two different approaches have been adopted to simulate auditory perception. One approach considers simplified assumptions regarding several auditory processing stages and models the effective transformations of these stages (e.g. Florentine et al., 1999; Jepsen et al., 2008; Moore et al., 2016). These models can be termed model observers, or heuristic observers (Geisler, 2011). Although heuristic observers are generally faster to compute and can predict some of the results from human perceptual tasks, the lack of physiological detail represents a disadvantage for using these models to relate different types and degrees of peripheral cellular damage to perception. A second approach utilises detailed models of the auditory periphery that can account for accurate neuronal representations of the AN, using an optimal combination of the information available in the nerve (e.g. Heinz et al., 2001a,b; Colburn et al., 2003; Lindahl et al., 2019). With this second modelling approach, using an ideal observer to optimally combine information at the level of the AN usually requires a priori knowledge of which segments to analyse in the AN response (i.e., a particular time window or one particular metric versus another one). This forces the need to identify or guess the task-relevant information contained in the stimuli. The ideal observer model typically leads to a model that performs much better than the real human observer. Then, by evaluating across some experimental parameter, it can be deducted that both the real and the ideal observers make use of similar properties or cues in the data, but the real observer is suboptimal or limited by usually unknown factors (Geisler, 2011).

For both modelling approaches, the derived observer is inherently limited to the specific task. Even though some models may be successfully transferred from one task to another (Moore and Glasberg, 1996; Florentine et al., 1999), potential key information may be still excluded or not considered in the new task. Furthermore, identifying the relevant cues of the stimulus are often impossible to derive for real-world tasks.

This has led to a third novel approach, that is, using a DNN model that will in extension to the AN model to replace the ideal observer (Kell et al., 2018; Haro et al., 2020; Saddler et al., 2021; Francl and McDermott, 2022). DNN models have shown to reach human performance for various visual tasks (Golan et al., 2020), and they have been shown to obtain constraints and limits similar to humans when trained on natural stimuli (Francl and McDermott, 2022). The deep layering structure of the DNN allows the model to learn intermediate abstractions and combine these optimally (Lecun et al., 2015). Furthermore, even when trained for a specific task, the DNN might learn abstract representations that are useful for other tasks. This is exemplified by the general aspect of transfer learning (Tan et al., 2018). When using DNN models on audio stimuli, it was shown that some of the learnt features are similar to some properties of the peripheral auditory system (Luo and Mesgarani, 2019).

The human auditory system is a result of evolution and thus an optimisation towards the environment in which humans exist. Despite the optimisation of a DNN model being purely artificial, the concept is similar. This makes the DNN models practical for replacing the ideal observer, as the models can be considered as an estimation of the ideal observer with less constraints on the information, without sacrificing the generality and applicability of the AN-model.

In this study, we have used a state-of-the-art model of the AN together with a convolution neural network (CNN) to investigate the potential use of gap detection to estimate CS in human listeners.

Lobarinas et al. showed elevated GDTs for chinchillas with substantial IHC-loss. Zeng et al. showed elevated GDTs in human patients with neuropathy. This may indicate that gap detection thresholds is sensitive to CS. We build and evaluated a computer modelling framework to test this hypothesis.

2 Method

2.1 Modelling framework

The general pipeline is described in fig. 1. Each block represent one general stage in the model framework, starting from the acoustic stimulus to the final decision variable. The inclusion of a physiologically-plausible AN-model allows for sufficient control over key parameters of the auditory periphery.

2.1.1 *Stimuli*

All stimuli were generated digitally with a sampling frequency of 100 kHz and saved as 64 bit floating-point encoded Resource Interchange File Format (RIFF) files. Stimuli was normalised by its respective root mean square (RMS)

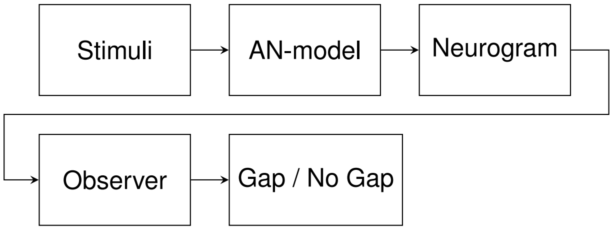

Fig. 1. Full feed-forward model structure, from stimulus to gap/no-gap label. Stimuli are the generated sounds. The AN model simulates the neural activity in response to the stimuli. The neurograms are constructed based on the output of the AN model, and then parsed to the observer (e.g. the DNN model), which outputs the probability of ”gap” and ”no gap”.

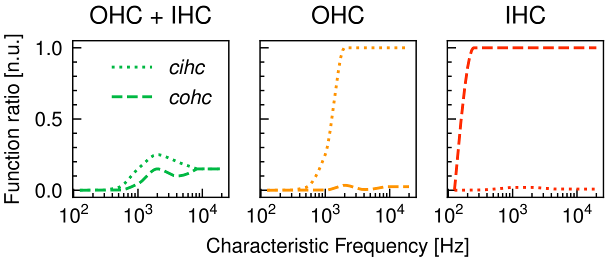

Fig. 2. Control parameters*cihc* and *cohc* for each of the elevated auditory threshold conditions inflicted by either a combination of OHC and IHC dysfunction (green) or either IHC (red) or OHC (orange) dysfunction only, respectively.

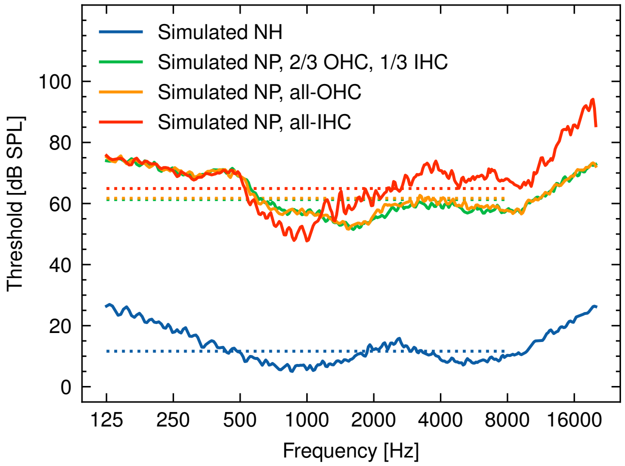

Fig. 3. Simulated model threshold for the NH and the NP hearing threshold condition from (Zeng et al., 2005). The NP threshold are simulated using three different combinations of IHC and OHC dysfunction. The solid lines show the simulated model threshold for the NH, $^2/_3$-OHC $^1/_3$-IHC, all-IHC and all-OHC conditions in blue, green, red and orange, respectively. The dashed lines show the mean thresholds from 125 to $8000\,\mathrm{Hz}$ for each simulated threshold .

value and calibrated to distinctive sound pressure levels (SPLs) from 20 to 100 dB SPL in steps of 5 dB.

The model framework does not impose any constraints on the presented stimulus. However, when using a DNN-model, the stimuli for training should follow the same considerations as when training any deep learning model.

For a more detailed description of the stimuli used for training the DNN model, see section 2.5.1.

2.1.2 AN model

The model of the AN by (Bruce et al., 2018) was used to simulate the AN response to sound. The humanised version of the model was used following similar parameters as used in previous studies (Encina-Llamas et al., 2019, 2021; Lindahl et al., 2019). In brief, the model simulated 200 characteristic frequencys (CFs) ranging from 125 to 20 000 Hz. A total number of 32 000 AN fibres, non-uniformly distributed along the CF axis (Spoendlin and Schrott, 1989), were simulated for the NH case. The distribution of AN fibres types were selected as 61 % of high spontaneous rate (SR) (HSR) fibres, 23 % of medium SR (MSR) fibres and 16 % of low SR (LSR) (Liberman, 1978).

For the CS simulations, the total number of fibres was reduced uniformly across CF and fibre type by a percentage loss from 20 to 80 % in steps of 20 %. For each CF, all the individually simulated fibres ascribed to that CF were added representing a summed response as function of time. The time-bin resolution is limited by the sampling frequency of the model (100 kHz). This time resolution was later reduced when constructing the neurogram (see section 2.1.3).

Two auditory thresholds was simulated, matching the two groups, NH and neuropathy (NP), from (Zeng et al., 2005). For the NH simulations, no impairment of the auditory periphery was implied. For the NP simulations, the MATLAB-function ‘fitaudiogram2’ developed by (Zilany et al., 2009) was used to impose IHC and OHC dysfunction to account for the elevated hearing thresholds. We used three dysfunction conditions with different combinations of IHC and OHC dysfunction, namely $^2/_3$-OHC $^1/_3$-IHC, all-IHC and all-OHC. Figure 2 shows the dysfunction control parameters for the three dysfunction conditions across CF. The control parameters *cihc* and *cohc* modify the IHC transduction and the OHC gain in the model for each simulated CF. The control parameters range from 0 to 1 ; 0 representing complete dysfunction and 1 representing healthy cells. Thus, the NH case was simulated with a constant value of 1 for both control parameters of IHC and OHC function. For the all-IHC and all-OHC conditions, the opposite control parameter is not completely 1 at all CF (middle and right panels in fig. 2). This is caused by the fitting process not being able to account for the full hearing threshold elevation by adjusting only the IHC or the OHC parameter. Figure 3 shows the simulated model threshold for all four AN model conditions using the method presented in (Encina-Llamas et al., 2018). For each CF, the method simulates a fixed number of 100 AN fibres for a pure-tone and silence to obtain two AN rate distribution (sound-driven activity vs spontaneous activity). A two-sample permutation test for equality of the means (Ernst and Bülthoff, 2004; Fisher, 1935) with 10000 permutation and a significant level of $1\,\%$ was used to find the model threshold.

2.1.3 Neurogram

The output of the AN model can be represented by a matrix of dimension $T \times F \times N$, with F being the number of simulated CFs, N the total number of time steps and T the AN fibre type. For a $1\,\mathrm{s}$ simulation, this results in $F = 200$ and $N = 100000$ per fibre type, which would require an immense amount of computational power for the

DNN model, following this block. In animal physiological recordings, neuronal responses are typically shown in the form of a peri-stimulus time histogram (PSTH). A PSTH implies binning together the spike counts within a time window. Previous studies that have used a similar setup with a model of the AN preceding a DNN model applied a temporal window to reduce the input along the temporal axis. Various window lengths have been used ranging from $0.05\,\mathrm{ms}$ in Saddler et al. (2021), $0.125\,\mathrm{ms}$ in Francl and McDermott (2022) to $8\,\mathrm{ms}$ in Haro et al. (2020). These studies used the combined setup to simulate pitch detection tasks (Saddler et al., 2021), binaural processing tasks (Francl and McDermott, 2022) and a task of speech recognition of digits (Haro et al., 2020).

The AN-model simulates the response of all the three AN fibre types independently (i.e., LSR-, MSR- and HSR). The framework allows for generating neurograms (S) with and without such third dimension of fibre type (see eq. (1)).

$$S \in \mathbb{R}^{1\times 200\times 800} \vee S \in \mathbb{R}^{3\times 200\times 800}\,. \tag{1}$$

In the case of summing across fibre type, the AN representation was a 2-D input matrix of size $F\times N$. On the other hand, if fibre type would not have been summed, the input matrix would be a 3-D matrix of the form $T \times F \times N$. Hence, the general model framework allows for either single channel or multi-channel neurogram, similar to the grey-scale or red, green and blue (RGB) colour image representation used as inputs to DNN models.

In the present study, all the contributions of all three fibre types were added to obtain a summed neural response per CF. This was decided based on considerations discussed in section A.4. A window size of $1\,\mathrm{ms}$ was used in the time binning, with no overlap.

2.1.4 Observer

This part of the framework was designed such that any model, in theory, could be evaluated. The requirement was simply a model that would compute a decision variable based on a single neurogram, ideally in the range $[0, 1]$,

$$\text{model} : \mathbb{R}^{T\times F\times N} \to \mathbb{R}\,. \tag{2}$$

The present study considered two types of models, a DNN based model (section 2.3) and a neurometric (NM) model (section 2.4).

If the model is capable of adapting to training data, e.g. such as the DNN-model, the framework enables the selection of distinct training data to be used and efficient loading of training data in mini-batches. This is used to evaluate the output of a model architecture trained on different data, which allows to investigate the aspects of the model being exploited during its optimisation process.

2.2 Gap detection thresholds

In literature, most GDT measurements use a n-alternative forced choice (AFC) experiment paradigm, with a 1-up, 2-down paradigm. To get a comparable measure for the framework, a setup using multiple samples of a range of gap lengths was used. The output of the DNN model was computed for a number of simulations without a gap, enabling the derivation of an internal response distribution for the no gap (noise) condition. A similar internal response distribution for the gap condition (noise + signal) was computed for a number of different lengths, spanning a range from very short gaps expected to be undetected to sufficiently long gaps expected to be detected.

The output of the DNN model was computed for 15 individual simulations for the no gap condition, and 15 simulations for each gap length conditions in the range from 1 to $39\,\mathrm{ms}$, in steps of $1\,\mathrm{ms}$, accumulating to 600 samples per test condition.

The DNN model was trained as a multi-class classifier with two output classes, ”gap” and ”no gap”. The DNN model used softmax as activation function, thus the summed output would always amount to 1 and therefor the output for the class ”gap” was solely used in the computation of GDT.

To obtain a comparable threshold for each model, a psychometric function with location and scale was fitted to the direct output of the DNN model to determine whether a gap was detected or not. An optimal fit for a model should have the location l as the mean of the output for the no gap condition and scaled with s such that the maximum value of the fitted function is the mean of the longest simulated gap length, to ensure that this is detected. This method is comparable with that of an optimal detector, where the decision criterion is set to the average of a given noise and noise + signal distribution (e.g. Jones, 2016), and thus generalises to other detection tasks.

We used the following definition of a psychometric function with location l and scale s,

$$\mathrm{psy}(x, \alpha, \beta, l, s) = l + s\frac{1}{1 + \exp\left(-\frac{x-\alpha}{\beta}\right)}\,, \tag{3}$$

where x is the gap length, α the centre position and β the slope of the psychometric function.

Another method used was to fit the psychometric function to the ratio of correct answers. The threshold of the decision variable to be correct was set to 0.5, indicating a more conservative approach. In this case, the placement of the no gap condition at $0\,\mathrm{ms}$ does not lead to a meaningful interpretation, hence it was excluded from the fitting process of the psychometric function. Furthermore, the location and scale was fixed by $l = 0$ and $s = 1$, as the ratio of correct answers should start at 0 for gap lengths where no gaps are detected, and converge to 1 for gap lengths where all gaps are detected. This simplifies eq. (3) to

$$\mathrm{psy}_{\text{correct ratio}}(x, \alpha, \beta) = \frac{1}{1 + \exp\left(-\frac{x-\alpha}{\beta}\right)}\,. \tag{4}$$

A third method was evaluated by computing an ideal decision criterion based on the model response. We observed that the DNN-model showed a bias towards higher output values when the level of the stimuli was low. Using fixed thresholds for all input levels is comparable with a yes/no task, while deriving an ideal threshold resembles more an AFC task. The ideal threshold was computed using the mean of the response to the no gap condition $\mathbb{E}\left[\mathrm{dv}_{\text{noise}}\right]$

and the mean of the response to the largest gap condition $\mathbb{E}\left[\mathrm{dv}_{\text{noise + signal}}\right]$ ($39\,\mathrm{ms}$),

$$\mathrm{dc}_{\text{ideal}} = \frac{1}{2}\left(\mathbb{E}\left[\mathrm{dv}_{\text{no gap}}\right] + \mathbb{E}\left[\mathrm{dv}_{\text{longest gap}}\right]\right) \ . \tag{5}$$

Using a 1-up, 2-down scheme, the GDT should indicate a gap length where the listener detects $70.7\,\%$ of correct responses (Levitt, 1971). This point relates to the psychometric function, as the function is an estimate of the expected ratio of correct responses. The threshold of the model was selected from deriving the intersection of eq. (4) at a given correct ratio p, leading to:

$$\mathrm{psy}_x(\alpha, \beta, p) = \alpha - \beta \log\left(\frac{1}{p} - 1\right) \ . \tag{6}$$

Note that this computation is independent of location and scale, as the placement along x only depends on parameters α and β

2.3 Deep neural network model

The deep learning modelling was implemented using the Python framework PyTorch (Paszke et al., 2019).

The model architecture was based on the general outline used by Kell et al. and modified by Haro et al.. It consisted of five convolutional blocks, followed by a fully-connected layer, a rectified linear unit (ReLU) activation layer, an output-layer and finally a softmax activation layer. Initial experiments in the present study indicated that model architectures with multi-dimensional kernels (e.g., kernel sizes larger than unity in the CF axis) outperformed human behaviour substantially and were insensitive to CS. This led to the choice of considering temporal-only kernels (i.e., a dimension of 1 in the CF axis).

The general structure is shown in fig. 4. For each convolutional-block, the number of kernels is dependent on a channel-factor, C. Thus, the total number of kernels for each layer is based on this factor. This is used for the convenience of having a single scalar for modifying the number of kernels as a hyper-parameter.

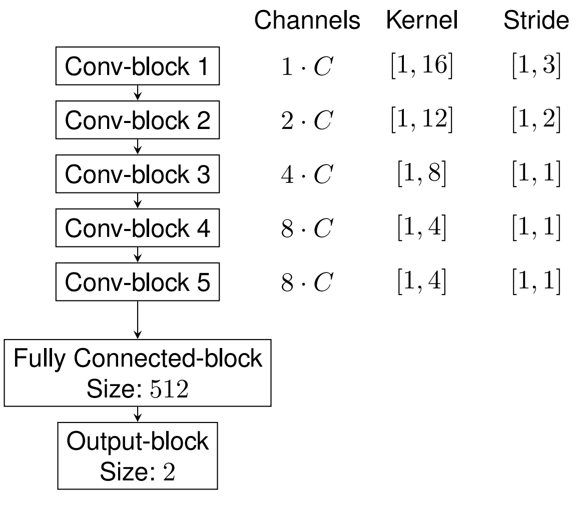

Fig. 4. Overview of the building blocks in the DNN-model architecture, with C as the channel-factor.

Each convolutional-block followed the structure shown in fig. 5. The input to the convolutional layer was zero-padded; hence the output retained the size (when stride=1). In the example of one-dimension, the input would be zero-padded due to the kernel-size k of the convolutional layer:

$$\text{padding} = \left(k - \left\lfloor\frac{k}{2}\right\rfloor - 1, \left\lfloor\frac{k}{2}\right\rfloor\right) \ . \tag{7}$$

In case of an uneven kernel size, the sides were padded equally by $\left\lfloor\frac{k}{2}\right\rfloor - 1$. For even kernel sizes, the left side were padded by 1 less than the right side. This is slightly different from the implementation of *same*-padding in the PyTorch Conv2d-class, and rather follows the method used in the deep-learning framework in TensorFlow (Abadi et al., 2016). The padding described above was applied for multi-dimensional kernels as well. In the present study, all convolution operations were using the 2-dimensional implementation in PyTorch. The kernel-sizes with one dimension equal to 1 were equivalent to a 1-dimensional convolution, but performed on each CF index.

The convolution layers used a variable number of channels, kernel size and stride, based on the hyper-parameters of the model (fig. 4). The layers also included bias, which for the actual implementation resulted in the following computation for convolving the layer weights W with the input X and adding the bias B.

$$Y_j = B_j + \sum_{k=0}^{C_{\text{in}}-1} W_{j,k} \star X_{i,k} \ , \tag{8}$$

with C_{in} being the number of input channels, j denoting the output channel index and $\star$ is the 2-D cross-correlation operator. The learnable weight matrix was thus of size $\mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k_0 \times k_1}$, for a convolutional layer with kernel size (k_0, k_1). The bias B is a single scalar for each output channel, thus $\mathbb{R}^{C_{\text{out}}}$.

The activation function used for the convolutional blocks was the ReLU function. This is the simplest rectifying activation function. Further optimisation by evaluating variants of the ReLU was not considered. Normalisation was done using local response normalisation (Krizhevsky et al., 2012), with fixed parameters of size $n = 5$, multiplicative factor $\alpha = 1e - 3$ and exponent $\beta = 0.75$. Pooling was done using average pooling, with a 2-dimensional kernel of size $(1, 3)$ (i.e., pooling was done in the time domain only) because the convolutional kernels only included the temporal dimension.

A fully connected block was placed after the series of convolution blocks. The output of the last convolution block was flattened to enforce a 1-dimensional representation of the 2-dimensional operation. Then followed a fully connected layer, a dropout layer and an activation layer, which for all cases was ReLU. This was decided based on previous evidence showing that the ReLU function is prominent in speech recognition tasks (Zeiler et al., 2013).

The output block consisted of a block with 2 neurons, one for each output label, followed by a softmax activation layer.

All DNN models were trained as a multi-label classification problem, with 2 labels: ”gap” and ”no gap”. We used

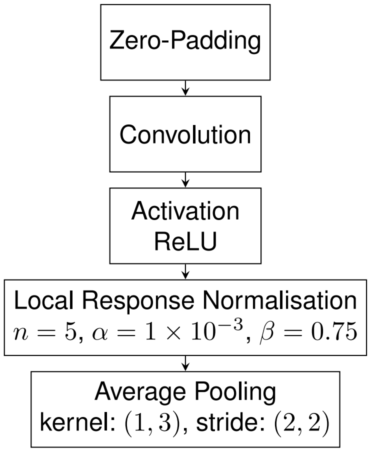
Zero-Padding
Convolution
Activation
ReLU
Local Response Normalisation
$n = 5$, $\alpha = 1 \times 10^{-3}$, $\beta = 0.75$
Average Pooling
kernel: $(1, 3)$, stride: $(2, 2)$

Fig. 5. Overview of a single convolutional block. The zero-padding is added such that the output of the convolution is shaped similarly to the input.

binary cross entropy for computing the loss, as implemented in PyTorch, where the loss for each label l_n is computed by

$$\begin{aligned} l_n =& y_n \cdot \max\left(\log(x_n), -100\right) & (9)\\ &+ (1 - y_n) \max\left(\log(1 - x_n), -100\right) \ , & (10) \end{aligned}$$

with the log-functions clamped by -100 to avoid $-\infty$. The mean of the label losses is used as the total loss $\mathcal{L}$ for the given sample,

$$\mathcal{L}(x, y) = \frac{1}{N} \sum_{n=0}^{N-1} l_n \,. \tag{11}$$

The training was performed using mini-batches of 64 samples. We used the ADAM optimiser (Kingma and Ba, 2015) with a weight decay of 0.5 and a learning rate of 0.0001, as implemented in FastAI (Howard and Gugger, 2020). Here, the applied regularisation is the product of the weight decay parameter and the learning rate.

A common method to extend the dataset used for training deep learning models is data augmentation. The simulated neurograms were 800 samples long, but the input of the DNN-model was chosen to support only 700 samples. This allowed for data augmentation by randomly shifting the input by up to 50 samples, and still retain the onset and the full length of the marker.

2.4 Neurometric model

To have an ideal equivalent model for comparison with the DNN model, we developed a simple model based on the immediate available information in the neurograms. The model used the change in rate with the *a priori* knowledge of position of the gap in the stimuli. This is a rather simple concept, exploiting rate as the primary information source, without extracting additional metrics such as synchrony or other second order metrics.

The neural metric model used the mean rate across CF, simply by summing all CFs for each time bin. Using the definition of the neurogram in eq. (1), the summed representation is defined by

$$s_{\mathrm{NM}}(n) = \sum_{t=0}^{1 \vee 3} \sum_{f=0}^{200} S(t, f, n) \,. \tag{12}$$

Note that for both neurogram definitions with either all fibre types summed or as different channels, the 1-dimensional reduction mentioned before is identical.

To reduce the fluctuation of the rate in each bin, the neural metric model used a moving average window, implemented by convolving the 1-D neurogram with a kernel of size K, with the kernel weights w_k defined by

$$w_k = \frac{1}{K} \,, \tag{13}$$

using the same definition as in eq. (8), but without bias. In the 1-dimensional case, this simplifies to

$$y = w \star s_{\mathrm{NM}} \,. \tag{14}$$

The forward difference of the averaged neurogram was then computed, and the resulting output was truncated from 200 to $600\,\mathrm{ms}$. The difference between the maximum and minimum forward difference was computed as the decision variable,

$$\mathrm{dv} = \max\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right) - \min\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right) \,, \tag{15}$$

where

$$\frac{\mathrm{d}y}{\mathrm{d}t} \approx \frac{\Delta y}{\Delta t} \,, \tag{16}$$

with $\Delta t = 1\,\mathrm{ms}$. To use and compare the model within the given framework, the output values for each experiment parameter were normalised by the no-gap condition and the longest simulated gap length condition. This gave a decision variable in the range $[0, 1]$. This normalisation requires an extra step in terms of the general structure of the modelling framework proposed.

2.5 Datasets

The full set of generated stimuli was divided into two main groups; training and testing. The DNN-model was only evaluated on the test data; i.e., the weights of the model were not updated in accordance to the loss associated with the testing samples.

2.5.1 *Training dataset*

The training data consisted of the two main groups of data. The first group of training data used the broadband Gaussian noise stimuli used in (Zeng et al., 2005), but with the gaps placed off-centre in four different times from the onset; 250, 300, 450 and $500\,\mathrm{ms}$. This subset contained 300 gap samples and 300 no-gap samples. The 75 gap lengths were selected by first computing 75 equally spaced lengths from 1 to 41, and later converting them using the following equation:

$$y = 10^{\frac{x}{20}} \,, \tag{17}$$

The log-spaced gap lengths were finally mapped to the range between 1 to $50\,\mathrm{ms}$ as:

$$y_{\mathrm{mapped}} = 1 + 49 \frac{y - \min(y)}{\max(y) - \min(y)} \,. \tag{18}$$

This emphasised gap lengths of short duration. The BBN training set was simulated at 11 sound levels, normalised by the total RMS value of the SPL of the given sample. The sound levels ranged from 20 to $70\,\mathrm{dB\,SPL}$, in steps of $5\,\mathrm{dB\,SPL}$.

The second group of training data used more naturalistic stimuli, namely multi-babble noise, in which a gap was added. Gap duration and position was sampled uniformly for 400 samples of each babble noise condition (see section C for a full description). Each sample was presented both with and without gap to avoid any sample specific cues for each sample. Each n-talker condition was simulated at 10 sound levels; 20, 25, 30, 40, 50, 60, 70, 80, 90 and 100 dB SPL. The number of talkers ranged from 3 to 8 , amounting to 48 000 simulations.

All training and testing simulations had a total duration of 800 ms. The noise stimuli were preceded by a period of 100 ms of silence. The duration of the noise stimuli was set to 500 ms, and was then followed by 200 ms of silence. All edges of the stimuli (i.e., the onset and offset of the noise and the gap) were ramped by a Hann-function of 5 ms. The first half of the window was used for onsets, while the second half was used for offsets, thus the ramp duration was 2.5 ms at all instances. All ramps were centred at the onset/offset time. See fig. 11 for an illustration of the applied envelope to the noise stimuli.

2.5.2 *Testing*

The testing dataset were simulations obtained using the same stimuli as used in (Zeng et al., 2005) for gap detection. These simulations was simulated as described for various degrees of CS and IHC and OHC dysfunction as described in section 2.1.2. The GDT obtained by Zeng et al. (2005) was presented in $\mathrm{sensationlevel(SL)}$, thus we converted the presentation levels by using the simulated pure-tone average threshold (PTA) for each of the auditory threshold conditions, shown in fig. 3. We trained four DNN models with the same hyper parameters, but randomly initialised weights. The GDT were obtained for each model, thus considering each model as an independent listener.

We found simulating an n-AFC to introduce additional variance when simulating a simple equivalent setup with normal distributions to model the internal response and thus we used the methods described in section 2.2.

We used psychometric functions fitted to percentage correct using an estimated ideal decision criterion, based on the considerations described in section A.5.

3 RESULTS

3.1 Sensitivity to CS

Figure 6 shows the effect of CS using the DNN model and the NM model, compared to the reference results from Zeng et al. (2005). The leftmost panel shows the group-averaged results from Zeng et al. (2005) for the 7 NH listeners and the 20 NP listeners. The human results show a clear difference in GDT between the groups, particularly at supra-threshold levels. At presentation levels near hearing threshold (5 dB SL), the mean GDT for the two groups were similar. For increasing presentation level, the GDTs for the NH group quickly reduced as a function of presentation level, flattening out at higher presentation levels beyond 40 dB SL. In contrast, the GDTs for the NP group decreased more shallowly and flattened out between around 13 ms until 40 dB SL, where they decreased further.

The simulated GDTs as a function of presentation level are shown in the middle panel. The results of the DNN model showed a difference between the simulated GDTs with 80 % of CS versus NH. Similar to the human data, the simulated GDTs decreased rapidly with increasing presentation level flattening out at supra-threshold levels. In contrast to the human data, the simulations with CS did not lead to a shallower decrease of GDT with increasing level. It did show a similar decrease at 40 dB SL with the human data. The simulated GDTs in the DNN model were slightly higher than the human threshold for the NH condition. The CS condition was simulated without imposing dysfunction to either IHCs and OHCs. Thus, the only change in the AN model between the two results in the DNN model is the number of AN fibres, which were evenly reduced for all fibre type and across all CFs.

The simulated results using the NM model with a window length of 15 ms are shown in the rightmost panel. The simulated GDTs showed no difference between the NH condition and the CS condition. The reduction of GDTs with increasing presentation level could also be observed using this model.

3.2 Sensitivity to hair cell dysfunction

Figure 7 shows the DNN model results from fig. 6 along with the simulations using the combined $^2/_3$ OHC and $^1/_3$ IHC dysfunction to account for the elevated auditory threshold of the NP group in (Zeng et al., 2005). The GDT is slightly lower for the simulated NP group at supra-threshold for both with and without CS. The NP show a steeper decay in GDT as a function of level, most notably by the NP results without CS which shows a GDT of 3 ms at a presentation level of about 12 dB SL. At the highest presentation level, the GDTs for the NP condition starts to increase while the CS condition show some fluctuation, ultimately reducing the difference between the two conditions at the highest presentation level. For both simulated groups without CS, the GDTs decays to a minimum level before increasing again. Notably at 28 dB SL for the NH condition and at 18 dB SL for the NP condition.

3.3 Sensitivity to specific hair cell dysfunction

Figure 8 shows the effect of independent hair cell dysfunctions, primarily by either IHC or OHC dysfunction. The all-IHC results show a large increase in GDT before lowering at 20 dB SL. It doesn't appear to reach a plateau similar to the other simulation conditions, however simulating higher presentation levels (above 100 dB SPL) would be unrealistic, given that this pressure level would be highly uncomfortable and damaging for the test subject. The all-OHC dysfunction case shows an even more rapid decay, compared to the other conditions, and generally obtains the shortest GDTs across all conditions. A dip in GDT is seen right after the initial decay, when following the GDT for the all-OHC condition from the lowest presentation level and onward. This dip is similar to the one observed for the other simulation conditions without CS, excluding the all-IHC dysfunction. The increase towards maximum presentation level is less clear compared to the result for the combined OHC and IHC dysfunction in fig. 7. The simulations didn't include any degree of CS.

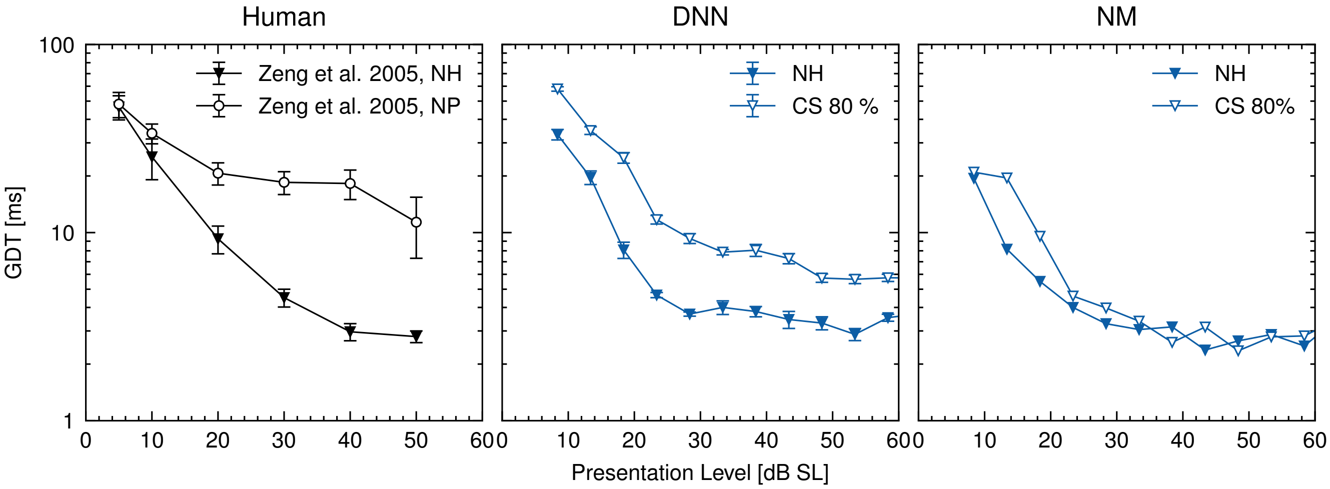

Fig. 6. Effect of CS on simulated GDTs as a function of presentation level (in dB SL for the DNN (middle) and NM (right) models compared to the reference results in humans from Zeng et al. (2005) (left). Downward solid triangles show GDTs for the NH results. Open circles (left panel) show the human NP results in Zeng et al. (2005). Downward open triangles (middle and right panels) show simulated GDTs for models including $80\,\%$ of CS. Error-bars indicate the ± SE. For the DNN-model, this was computed from 4 identical model architectures, while the behavioural results included 7 NH listeners and 20 listeners with NP.

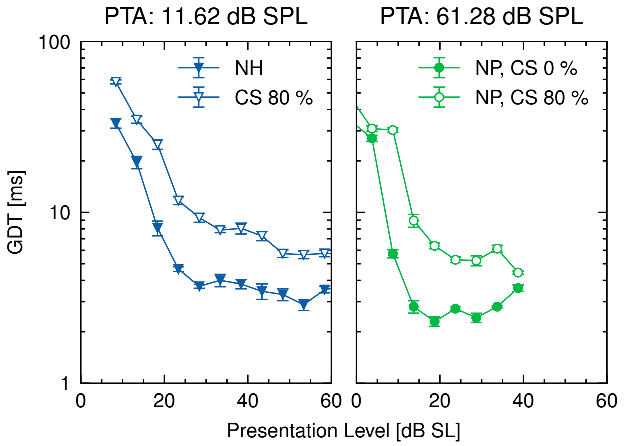

Fig. 7. Effect of combined OHC and IHC dysfunction (green) as function of presentation level (dB SL) in combination with the simulated NH GDT (blue) copied from the middle panel of fig. 6. Open symbols show the results for the simulations with $80\,\%$ CS.

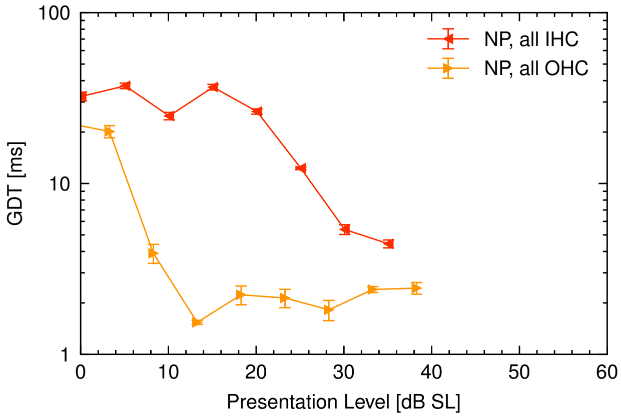

Fig. 8. Comparison of GDT by simulating all-IHC (red) and all-OHC (orange) dysfunction as a function of presentation level (dB SL).

4 Discussion

4.1 General overview

We developed a model framework for evaluating behavioural tests that could be sensitive to hearing impairments not reflected by an audiogram. Specifically, we evaluated this modelling framework on a gap detection task for both a simplistic NM model and a DNN model following a detailed model of the AN.

We trained the DNN model on NH simulations of $500\,\mathrm{ms}$ long gap carriers (noise stimuli containing a gap) on combinations of babble noise and BBN. We found that it was needed to include more naturalistic stimuli in the training data (see section A.2) to achieve human-like behaviour, as similarly found in previous studies (Saddler et al., 2021; Francl and McDermott, 2022).

We then evaluated the DNN model on a test set with NH neurograms and neurograms with various impairments inflicted by either a combination of IHC and OHC dysfunction and separately, with and without CS. Both the DNN and the NM model resulted in near human-like performance in the normal hearing case. While the DNN model showed elevated GDTs for a high degree of CS, the NM model showed little to no difference related to CS. Furthermore, the DNN model showed equal or lowered GDTs when the simulated auditory threshold shift was imposed majorly by OHC dysfunction.

4.2 Sensitivity to CS

In fig. 6 we reproduced the data for the gap detection task in (Zeng et al., 2005) for the two groups, NH and NP. Starting with the NH, it can be observed that the GDT is largely affected by the sound presentation level. At higher levels, the GDT seems to reach a plateau. This is in agreement with the findings in studies evaluating the effect of level on gap detection (Irwin et al., 1981; Buus and Florentine, 1985; Moore et al., 1992). Moore et al. (1992) described a comparable effect when measuring GDT with pure-tone carriers at different frequencies. They found that the GDT reached an asymptote at 55 dB SPL for all tested frequencies between 100 to 2000 Hz.

A clear change in GDT is seen between the NH and NP groups (left panel in fig. 6). The GDTs in the NP group shows a larger variance. This is consistent with other studies reporting larger GDT variance in impaired listeners compared to NH control (Irwin et al., 1981; Fitzgibbons and Wightman, 1982; Buus and Florentine, 1985; Moore et al., 1992). This may indicate that the underlying mechanisms are affected in different degrees. Furthermore, this difference in GDT wasn't associated with the differences in auditory hearing thresholds for the groups.

The mid panel of fig. 6 shows the modelled GDT for NH simulations with and without CS. The GDT for the CS condition was elevated, but not as much as for the NP group from Zeng et al. (2005). The human results indicated a common limit at low presentation levels for both groups. This similar shared point of limitation was not observed in the model results with and without CS. Rather, the results with CS seems to be shifted by presentation level. This might be explained by various factors further discussed in section 4.5.

The NM model seems to achieve GDTs closer to the human results for the NH condition. However, there is no clear elevation of the GDTs for the CS case. For the other conditions of hearing impairment, similar results were obtained, thus the NM indicates no peripheral impairment affecting the GDT, contrary to the results presented by Lobarinas et al. (2020). At low presentation levels, a shift of around 5 dB was observed, similar to the shift observed for the DNN model.

4.3 Sensitivity to hair cell dysfunction

The effect of hearing impairment on the DNN model was simulated by fitting the mean audiogram of the NP group from Zeng et al. (2005) to the AN model, both with and without CS. The results are shown in fig. 7. The simulated results showed an overall reduction of the GDT in the hair cell impairment condition, both with and without CS. This is opposite to the expected effect and the experimental results shown in literature. The plot obtained for the combined IHC and OHC dysfunction, with and without CS, closely resembles two individuals measured by (Buus and Florentine, 1985). One subject showed GDT comparable to the NH group, while another subject obtained a minimum GDT at around 7 ms. Furthermore, the latter subject had considerably lower auditory thresholds than the former. As the presentation level is increased beyond 70 dB SPL, the DNN model generally produced higher GDTs. A similar effect was observed in several human studies with noise carriers of different spectral width (Buus and Florentine, 1985; Moore et al., 1992).

Studies on the effect of stimulus frequency on GDT have shown higher thresholds for lower centre frequencies (Shailer and Moore, 1983; Florentine et al., 1999). It has been proposed that the "ringing" of the basilar membrane (BM) response could contribute to this added constraint. The BM does not instantaneously stop moving at the offset of a stimulus. For higher frequencies, this effect may be minimal as the vibration of the BM decays more quickly. By modelling the BM as a filter bank with time-varying level-dependent filters to model the effect of OHC gain on BM motion, the low CF filters "ring" for a longer duration due to their broader frequency tuning. This effect is easily observed in the neurograms at the offset of the BBN carrier. Two effects depending on stimulus level were visible. Firstly, toward low levels, the onset following the gap was weaker while the firing did not completely stop during the gap. Both properties reduced the distinctiveness of the onset following the gap. Secondly, towards high levels, the tail following the onset of the gap increased in duration, potentially obfuscating the onset following the gap. The combination of these two effects could explain how the GDT increased for very low and very high presentation levels.

When OHCs are impaired, the sharp tuning of the BM is reduced. In line with the above, the neural firing following the onset of the gap is reduced. This would explain the reduced GDT for the combined IHC and OHC simulations and all-OHC simulations without CS (figs. 7 and 8). The reduced activity was clearly observed in the neurograms.

4.4 Effect of either IHC or OHC dysfunction

As shown in fig. 8, GDTs were substantially increased for the all-IHC dysfunction simulation. The DNN model showed a large bias towards "gap" up until about 20 dB SL, similar to the response for the NH simulations at low levels. This bias is discussed in further depth in section 4.6.5. Furthermore, the DNN model did not achieve above 70 % correct responses at 10 dB SL. The psychometric functions fitted for the correct responses for up to 20 dB SL were fitted to responses that did not reach 100 % correct within the gap length range included in the test set. Thus, the data basis for the thresholds is worse for these presentation levels. From about 20 dB SL, the GDT curvature started to resemble the NP with CS simulation in fig. 7 from about 10 dB SL. The effective threshold seems to be higher than the computed PTA, thus making it difficult to evaluate how the GDT compares if a lower hearing loss was simulated.

From the simulated results shown in fig. 7 and fig. 8, it seems clear that the GDTs of the DNN model were affected by impairment involving IHC dysfunction and AN fibre loss, but largely insensitive to threshold shifts due to OHC dysfunction. On the contrary, OHC dysfunction resulted in lowered GDTs for the DNN model. The reduction of GDTs has not been observed in any hearing impaired human data, which only showed equally good or worse GDT.

The combination ratio of 2/3-OHC and 1/3-IHC dysfunction was based on the suggestions stated by (Lopez-Poveda and Johannesen, 2012). However, within their results, a large variability exists. This might indicate differences in the combination of IHC and OHC dysfunction in different individual listeners and could additionally be an artefact due to using forward masking paradigms to estimate cochlear compression (Lindahl et al., 2019). The relation between the degree of IHC dysfunction onto GDT is not clear from our results. To better understand the effect of IHC dysfunction in this relation, simulations for various combinations of IHC and OHC dysfunction would be valuable. In comparison, the degree of CS affected the GDT exponentially at presentation levels above 25 dB SL. The same analysis couldn't be evaluated for the degree of IHC dysfunction due to the missing data points. To further understand the effect of IHC dysfunction, different auditory thresholds solely inflicted by IHC dysfunction, using the *cihc* parameter should evaluated.

4.5 Auditory thresholds

We computed the auditory thresholds of the model framework based on the output of the AN model with a limited set of fibres. This method is fast and it's favourable in terms of the DNN model only being trained for a single task. For the setup of the DNN model, where the output is the classification of either gap or no gap, it's not possible to determine a direct threshold of a given sound. This poses a problem when comparing the results of the model with human results obtained in SL with no reference for the auditory threshold of the given stimuli. The method we used for computing the auditory threshold isn't sensitive to CS, however, CS would theoretically increase the auditory threshold, shown by (Oxenham, 2016). Lobarinas et al. (2020) also observed threshold shifts, though within what would be categorised as NH. As Lobarinas et al. (2020) didn't measure IHC dysfunction apart from complete loss, it's unclear whether the discrepancy with the theoretical caused by CS is due to the combination of information in Oxenham (2016) or undiscovered IHC dysfunction in addition to the IHC loss in the animals.

The GDT shown for NH and 80 % CS in fig. 6, indicates a upward shift for the CS condition along level. This might be explained by the auditory threshold shift due to CS, as described above, but it gets more complicated when we compare with the simulations with combined IHC and OHC dysfunction shown in fig. 7. Here, the CS condition seems shifted. Additionaly, this might as be an discrepancy between our method of computing SL and the method used by (Zeng et al., 2005) as they used the auditory threshold for the carrier with no gap, while we used the PTA as reference.

4.6 Framework details

4.6.1 Spikeogram parameters

In (Zeng et al., 2005) it's suggested that temporal processing depends on a synchronised spike code for working optimally. However, the fine structure of the sound was effectively removed as we transformed the AN simulations to neurograms. By using a bin width of 1 ms, we removed any frequency components of the spike code above 500 Hz. This information doesn't seem critical and even without, the DNN model learned features dependent on the redundancy of AN fibres. By reducing the bin width, we would remove less information but at the computational expense of a larger input. Phase locking is said to be a factor up to 1.5 kHz, thus an optimal bin width would be at least twice this frequency, i.e. at least 3 kHz. This might be a reason why the results are good for (Saddler et al., 2021) and (Francl and McDermott, 2022), as by using sampling frequencies at 20 kHz and 8 kHz respectively, allows the DNN model to gain advantage from said phase locking cues. We didn't evaluate this as changing the bin width also changes the temporal length of the kernels, making it difficult to compare.

4.6.2 Stride of convolutional kernels

The DNN model was implemented with stride for the longer kernels in first two convolution blocks. This is usually implemented to reduce the size of the feature space following the processing block, effectively down-sampling the input with learnable filters. Following the convolution layer we applied average-pooling, further reducing the size. This setup originated from the models with multi-dimensional kernels, discussed in section A.1.1, and was only changed in the spectral domain, thus that the kernels didn't stride in the spectral domain. The pooling layer still strided by 2 in both dimensions, though pooling kernel only averaged in the temporal domain. We therefore checked whether an effect was visible by setting the stride to 1 for all dimensions, with the conclusion that the major trend wasn't different, however the GDT across level was less fluctuating. Furthermore, the general thresholds was higher for the models with stride set to 1, thus the original results are kept.

4.6.3 Temporal only vs. spectral and temporal

We used the DNN model architecture with temporal-only kernels, as the DNN models with multi-dimensional kernels obtained GDTs substantially lower than the human reference, furthermore, these models wasn't substantially sensitive to CS with only ≈1 ms elevation (see section A.1.1). However, the temporal-only models didn't obtain as low GDTs as the NH human reference and showed a shift in threshold by simulating CS similar to the NP reference, though not as profound. A study using spectral-modulation filter banks was able to predict the trend of a frequency dependent gap detection task, but the model obtained significantly higher thresholds (Sanchez and Dau, 2016), suggesting an interplay of temporal and spectral information.

In the comparison between models with temporal and spectro-temporal kernels, we can deduct that the deep representation learned by the models with temporal-only kernels are highly dependent on the information provided by having more functioning AN-fibres, reflected by the increased GDT. However, there is evidence of spectro-temporal receptive-like fields in the auditory brain stem (Schönwiesner and Zatorre, 2009), thus it can't be excluded that along the auditory brain-stem, 2-D convolutional like receptors exist.

Gap detection is a simple task and the stimuli firing pattern in the AN is simple compared to more complex sounds, such as speech. The model is optimised towards the task and by the data presented. The cues needed for gap detection alone might be more simple than the cues used for speech recognition, thus enabling the spectro-temporal kernels to overfit to the task by learning shapes that otherwise would be used to encode the more complex shapes of speech.

The first layers of a CNN can learn to extract envelopes of the input. With the rectifying non-linearity, the layers can show resemblance with an envelope extractor in which the signal is squared and then low-passed. The following layers are not different, thus these are capable of learning to extract the modulation of the envelopes. Thus there might be similarities between the trained DNN and the heuristic models using modulation filter banks (Jepsen et al., 2008).

4.6.4 Smearing of the internal representation?

In (Zeng et al., 1999), it's discussed whether gap detection deficits in subjects with neuropathy can be explained by a simple smearing model. The model proposed by Zeng et al. (1999) is equivalent to convolving a window across the envelope of the stimuli. To test this strategy we employed the NM with a moving average window of 30 ms. The only thing changed in the NM model was the window length. However, the GDTs remained largely unchanged. The model didn't show substantial change of threshold for the CS conditions, leaving the conclusions of the validity of the NM unchanged. Thus, the window model suggested by Zeng et al. (1999) seem inadequate for explaining the elevated threshold.

4.6.5 Training data

For training we used both BBN and babble-noise in a non-even distribution between the two noise sources. The BBN included selectively lower gap lengths as a result from the logarithmic mapping (eq. (17)), resulting in 84 % of the gaps having a duration below 25 ms. In contrast, the babble noise dataset had approx. 25 % gaps of duration below 25 ms due to the uniformly sampling of gap lengths. When including the BBN in the training set, the GDT lowered, for all the test conditions. This could partially be by the added loss for lower thresholds, forcing the model to improve for lower gap lengths. In the initial trials, the models trained for BBN only showed a poorer ability to detect gaps with lengths above the seen 50 ms. This led to the use of both a uniformly distributed selection of gap lengths in the babble noise and to include the longer range of gap lengths from 0 to 100 ms.

The resemblance between the training and test data when including the BBN training data, may also be a contributing factor to the improved performance, being that the BBN used for the training set and test set shared the exact same parameters for bandwidth and ramping.

The selected DNN models for the results in fig. 6 had a bias towards classifying the input as ”gap” for the lower presentation levels. For 20 dB SPL, the mean output of the model to the NH no-gap samples was above 0.5, similarly to the mean output for the longest gap length. Similar to the analysis of model performance as a function of training iterations section A.3, we analysed the bias at 20 dB SPL, but found no trend across training. Notably, the mean of the outputs didn't go below 0.5.

Even though the effort to create a balanced training set, we found after generating the results presented that the BBN dataset had an imbalance in count of gap and no-gap samples. The simulations for BBN at dB SPL didn't include the no-gap samples, thus 300 no gap samples was missing from the total training dataset size of 54 300, i.e. the total number should have been 54 600. In term of number of samples, the error is non-significant, however, we deem the imbalance to as the main cause for the observed bias.

The BBN training set was only simulated for 20 to 70 dB SPL. Not visible from the plotted range in fig. 6, the computed thresholds had a tendency to increase for test levels higher than 70 dB SPL. A similar trend was observed for the CS simulations, but starting at a slightly lower presentation level and not always as profound. The trend is visible in fig. 9 for the tested channel factors. Furthermore, this effect is increasingly visible as the number of channels is reduced for the DNN model. Inspecting the weights of the first convolutional layer doesn't however show a clear tendency towards filters matched for certain levels, as the change in neural activity isn't linear, such a selectivity in bias and weights of the convolutional layer may also not be linear and hard to identify without knowing how the output of the first layer is processed in the later layers.

Three factors might therefore contribute to the DNN model ability to obtain low thresholds for all presentation levels far beyond hearing threshold, i.e. above 40 dB SPL; the presentation levels in the training dataset and the number of channels in the model. The third factor would be the effect also seen in human data.

4.6.6 Reference studies

We only evaluated the gap detection thresholds along one parameter, sound intensity. However, gap detection has been evaluated for a number of parameters as described in section B. To further strengthen the argument for the proposed model framework and the use of gap detection for assessing AN health, the results would benefit from adding human data evaluating GDT over other parameters than level.

Previous studies on gap detection saw a large effect of frequency on GDT (Shailer and Moore, 1983; Florentine et al., 1999). The evaluation of the model would benefit from including simulations of such studies to investigate the generalisation of the model behaviour. However, the reduced GDT for OHC dysfunction indicates that such behaviour frequency dependent behaviour is likely tied to the effect of the OHCs on the BM. Though, it doesn't hint at how the model would behave for narrow-band carriers.

Another parameter is carrier duration, thus using the study from Schneider and Hamstra (1999) as a reference. In the study, the GDTs two groups based on age was measured using gap carrier durations from 0.5 to 500 ms. At long durations, the two groups obtained similar GDT, but diverged as the duration was shortened, with the elderly group showing higher GDTs. As with the other studies on gap detection, the elderly or hearing impaired group showed larger variance. Adding this study would therefor benefit in two ways, firstly the model behaviour to shorter gap carriers is evaluated displaying whether the model generalises for other aspects of the same task, secondly, it would be enlightened whether the increased GDTs for shorter carrier durations is caused by age alone.

4.7 Model limitations and future directions

Deep learning as a tool for estimating perceptual models directly on detailed simulations of the AN have proven capable of reaching human-like performance. The present study supports this. While classic model observer and ideal observer approaches are limited by the simplification of the auditory periphery and the information integration, the deep learning approach introduces similar limits. The DNN model are limited by the task it's trained for and the data it's trained on. The implications of these limits might be concealed by the nature of deep learning models.

The present problem of defining a model capable of explaining human data in different behavioural tasks can be further developed in two directions.

One direction is to extend on the same approach used by Saddler et al. (2021); Francl and McDermott (2022), by sharing the learned weights across domains. This could be done by either using transfer learning or defining the models for multi-task learning. Transfer learning could enlighten what cues or mechanisms might be shared across behavioural tasks, e.g. by evaluating how the model would perform if only the last linear layers was retrained for the new task. Multi-task learning cold similarly show which features are shared by forcing the model to learn feature extractions that are optimal for both domains.

Another approach would be to tackle the problem from another angle. One aspect of the present study was to develop a model capable of evaluating behavioural tasks before performing the test on humans. The approach in the present paper relies on defining a new task with a similar output label, i.e. within the same domain. This limitation could be circumvented by training a model not limited by task but by procedure. Instead of outputting the probability of a certain pitch or gap and only present one stimulus at the time, the model could take n presentations and simply reply which of the n presentations was different. This would allow the model to generalise across behavioural tasks for different domains and thus only be constrained by the ability to define a domain specific measure as a n-AFC task.

Given that an above framework would function, this also opens the possibility of modelling how various rehabilitation strategies affect the performance in the behavioural tasks. A step further from finding a behavioural test suitable for detecting CS, is the need to find an effective strategy to reduce the impairing effects of the condition. As it's straightforward to apply digital sound processing to the stimuli before simulating the AN, the effect of various hearing aid processing strategies on various behavioural tests could be investigated.

5 Conclusion

The goal of this study was to both develop a general modelling framework for evaluating behavioural tests potentially sensitive to CS and specifically evaluate gap detection within this framework.

We showed that a model trained on natural babble noise was able to resemble the tendency of GDTs in NH human data showed sensitivity to CS reflected by elevated GDTs. Furthermore, we saw from simulating CS, IHC- and OHC dysfunction that CS was the primary contributor to elevated GDTs, IHC dysfunction seemed to contribute as well, while OHC dysfunction showed a direct opposite effect by lowering the GDTs.

Future work on extending the gap detection test references for the present model could serve to further strengthen the argument for using gap detection as a possible clinical test for assessing IHC function and CS.

6 Acknowledgements

All simulations and model training was conducted on the Technical University of Denmark (DTU) high performance computing (HPC) (DTU Computing Center, 2021) and this project would not have been possible without the sheer processing power available from the cluster.

7 Data availability

The full source code of the framework and the implemented DNN models is currently residing in a private repository, but will be made fully available at https://github.com/pjnr1/gapnet.

The complete data collection with stimuli, AN simulations, neurograms and model weights can be made available upon request.

Acronyms

ABR	auditory brainstem responses. 1
AFC	alternative forced choice. 4, 7, 12, 16
AM	amplitude modulation. 2
AN	auditory nerve. 1–4, 7–12, 16
BBN	broadband-noise. 1, 6, 8, 9, 11, 16
BM	basilar membrane. 9, 11
CF	characteristic frequency. 3–7, 9
CNN	convolution neural network. 2, 10
CS	cochlear synaptopathy. 1–3, 5, 7–12, 15–17
DNN	deep neural network. 1–12, 15–17
DPOAE	distortion product otoacoustic emission. 1
DTU	Technical University of Denmark. 12
EFR	envelope following responses. 1
GDT	gap detection threshold. 1, 2, 4, 5, 7–12, 15–17
HHL	hidden hearing loss. 1, 2
HPC	high performance computing. 12
HSR	high SR. 3, 4
IHC	inner hair cell. 1–3, 7–10, 12
LSR	low SR. 3, 4
MSR	medium SR. 3, 4
NH	normal hearing. 1, 3, 7–12, 15–17
NM	neurometric. 4, 7–9, 11
NP	neuropathy. 3, 7–10, 15, 17
OHC	outer hair cell. 1, 3, 7–12
PSTH	peri-stimulus time histogram. 4
PTA	pure-tone average threshold. 7, 9, 10
ReLU	rectified linear unit. 5, 6
RGB	red, green and blue. 4
RIFF	Resource Interchange File Format. 2
RMS	root mean square. 2, 6
SE	standard error. 8
SL	sensation level. 7–10
SPL	sound pressure level. 3, 6, 7, 9, 11, 16, 17
SR	spontaneous rate. 3, 13

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, page 265–283, USA, 2016. USENIX Association.

A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.

H. M. Bharadwaj, S. Masud, G. Mehraei, S. Verhulst, and B. G. Shinn-Cunningham. Individual differences reveal correlates of hidden hearing deficits. *Journal of Neuroscience*, 35(5):2161–2172, 2015.

N. Bramhall, E. F. Beach, B. Epp, C. G. Le Prell, E. A. Lopez-Poveda, C. J. Plack, R. Schaette, S. Verhulst, and B. Canlon. The search for noise-induced cochlear synaptopathy in humans: Mission impossible? *Hearing Research*, 377:88–103, 2019.

N. F. Bramhall, D. Konrad-Martin, G. P. McMillan, and S. E. Griest. Auditory Brainstem Response Altered in Humans With Noise Exposure Despite Normal Outer Hair Cell Function. *Ear & Hearing*, 38(1): e1–e12, 1 2017.

N. F. Bramhall, G. P. McMillan, and S. D. Kampel. Envelope following response measurements in young veterans are consistent with noise-induced cochlear synaptopathy. *Hearing Research*, 408:108310, 9 2021.

I. C. Bruce, Y. Erfani, and M. S. Zilany. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. *Hearing Research*, 360: 40–54, 3 2018.

S. Buus and M. Florentine. Gap Detection in Normal and Impaired Listeners: The Effect of Level and Frequency. In *Journal of Speech & Hearing Research*, volume 27, pages 159–179. Springer Berlin Heidelberg, Berlin, Heidelberg, 1985.

H. S. Colburn, L. H. Carney, and M. G. Heinz. Quantifying the information in auditory-nerve responses for level discrimination. *JARO - Journal of the Association for Research in Otolaryngology*, 4(3): 294–311, 9 2003.

W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann. ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology. *Audiology : official organ of the International Society of Audiology*, 40(3):148–157, 2001.

DTU Computing Center. DTU Computing Center resources. *Technical University of Denmark*, 2021.

G. Encina-Llamas, T. Dau, J. M. Harte, and B. Epp. A mouse model of the auditory nerve to study cochlear synaptopathy. In *41st Midwinter Meeting of the Association for Research in Otolaryngology, ARO 2018*, 2018.

G. Encina-Llamas, J. M. Harte, T. Dau, B. Shinn-Cunningham, and B. Epp. Investigating the Effect of Cochlear Synaptopathy on Envelope Following Responses Using a Model of the Auditory Nerve. *JARO - Journal of the Association for Research in Otolaryngology*, 382: 363–382, 2019.

G. Encina-Llamas, T. Dau, and B. Epp. On the use of envelope following responses to estimate peripheral level compression in the auditory system. *Scientific Reports*, 11(1):1–19, 2021.

M. O. Ernst and H. H. Bülthoff. Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162–169, 2004.

R. A. Fisher. *The design of experiments.* Oliver & Boyd, Oxford, England, 1935.

P. J. Fitzgibbons and F. L. Wightman. Gap detection in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 72(3):761–765, 1982.

M. Florentine, S. Buus, and W. Geng. Psychometric functions for gap detection in a yes–no procedure. *The Journal of the Acoustical Society of America*, 106(6):3512–3520, 1999.

C. Formby and T. G. Forrest. Detection Of Silent Temporal Gaps In Sinusoidal Markers. *Journal of the Acoustical Society of America*, 89(2): 830–837, 1991.

A. Francl and J. H. McDermott. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, 6(1):111–133, 2022.

A. Fulbright, C. Le Prell, S. Griffiths, and E. Lobarinas. Effects of Recreational Noise on Threshold and Suprathreshold Measures of Auditory Function. *Seminars in Hearing*, 38(04):298–318, 11 2017.

A. C. Furman, S. G. Kujawa, and M. C. Liberman. Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates. *Journal of Neurophysiology*, 110(3):577–586, 8 2013.

W. S. Geisler. Contributions of ideal observer theory to vision research. *Vision Research*, 51(7):771–781, 2011.

T. Golan, P. C. Raju, and N. Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(47):29330–29337, 2020.

H. Guest, K. J. Munro, G. Prendergast, S. Howe, and C. J. Plack. Tinnitus with a normal audiogram: Relation to noise exposure but no evidence for cochlear synaptopathy. *Hearing Research*, 344:265–274, 2 2017.

S. Haro, C. J. Smalt, G. A. Ciccarelli, and T. F. Quatieri. Deep Neural Network Model of Hearing-Impaired Speech-in-Noise Perception. *Frontiers in Neuroscience*, 14:588448, 12 2020.

N.-J. He, A. R. Horwitz, J. R. Dubno, and J. H. Mills. Psychometric functions for gap detection in noise measured from young and aged subjects. *The Journal of the Acoustical Society of America*, 106(2):966–978, 1999.

M. G. Heinz, H. S. Colburn, and L. H. Carney. Evaluating Auditory Performance Limits: I. One-Parameter Discrimination Using a Computational Model for the Auditory Nerve. *Neural Computation*, 13 (10):2273–2316, 10 2001a.

M. G. Heinz, H. S. Colburn, and L. H. Carney. Evaluating Auditory Performance Limits: II. One-Parameter Discrimination with Random-Level Variation. *Neural Computation*, 13(10):2317–2338, 2001b.

T. T. Hickman, C. Smalt, J. Bobrow, T. Quatieri, and M. C. Liberman. Blast-induced cochlear synaptopathy in chinchillas. *Scientific Reports*, 8(1):10740, 12 2018.

A. E. Hickox, E. Larsen, M. G. Heinz, L. Shinobu, and J. P. Whitton. Translational issues in cochlear synaptopathy. *Hearing Research*, 349: 164–171, 6 2017.

S. E. Hind, R. Haines-Bazrafshan, C. L. Benton, W. Brassington, B. Towle, and D. R. Moore. Prevalence of clinical referrals having hearing thresholds within normal limits. *International Journal of Audiology*, 50(10):708–716, 10 2011.

J. Howard and S. Gugger. Fastai: A layered api for deep learning. *Information (Switzerland)*, 11(2):1–26, 2020.

R. J. Irwin, L. K. Hinchcliff, and S. Kemp. Temporal acuity in normal and hearing-impaired listeners. *International Journal of Audiology*, 20 (3):234–243, 1981.

M. L. Jepsen, S. D. Ewert, and T. Dau. A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, 124(1):422–438, 7 2008.

I. H. Jones and V. O. Knudsen. Functional Tests of Hearing. *California and Western Medicine*, 23(9):1166, 9 1925.

P. R. Jones. A tutorial on cue combination and Signal Detection Theory: Using changes in sensitivity to evaluate how observers integrate sensory information. *Journal of Mathematical Psychology*, 73:117–139, 2016.

A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644, 2018.

D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

S. G. Kujawa and M. C. Liberman. Adding insult to injury: Cochlear nerve degeneration after ”temporary” noise-induced hearing loss. *Journal of Neuroscience*, 29(45):14077–14085, 2009.

G. Kumar, F. Amen, and D. Roy. Normal hearing tests: is a further appointment really necessary? *Journal of the Royal Society of Medicine*, 100(2):66–66, 2 2007.

Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

H. Levitt. Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2B):467–477, 1971.

M. C. Liberman. Auditory-nerve response from cats raised in a low-noise chamber. *The Journal of the Acoustical Society of America*, 63(2): 442–455, 1978.

M. C. Liberman, M. J. Epstein, S. S. Cleveland, H. Wang, and S. F. Maison. Toward a Differential Diagnosis of Hidden Hearing Loss in Humans. *PLOS ONE*, 11(9):e0162726, 9 2016.

H. W. Lin, A. C. Furman, S. G. Kujawa, and M. C. Liberman. Primary Neural Degeneration in the Guinea Pig Cochlea After Reversible Noise-Induced Threshold Shift. *Journal of the Association for Research in Otolaryngology*, 12(5):605–616, 10 2011.

J. C. T. Lindahl, G. Encina-Llamas, and B. Epp. Analysis of a forward masking paradigm proposed to estimate cochlear compression using an auditory nerve model and signal detection theory. *Proceedings of ISAAR 2019: Auditory Learning in Biological and Artificial Systems. 7th symposium on Auditory and Audiological Research*, 7(August), 2019.

L. Liu, H. Wang, L. Shi, A. Almuklass, T. He, S. Aiken, M. Bance, S. Yin, and J. Wang. Silent Damage of Noise on Cochlear Afferent Innervation in Guinea Pigs and the Impact on Temporal Processing. *PLoS ONE*, 7(11):e49550, 11 2012.

E. Lobarinas, R. Salvi, and D. Ding. Insensitivity of the audiogram to carboplatin induced inner hair cell loss in chinchillas. *Hearing Research*, 302:113–120, 8 2013.

E. Lobarinas, C. Spankovich, and C. G. Le Prell. Evidence of “hidden hearing loss” following noise exposures that produce robust TTS and ABR wave-I amplitude reductions. *Hearing Research*, 349:155–163, 6 2017.

E. Lobarinas, R. Salvi, and D. Ding. Gap Detection Deficits in Chinchillas with Selective Carboplatin-Induced Inner Hair Cell Loss. *Journal of the Association for Research in Otolaryngology*, 21(6):475–483, 12 2020.

E. A. Lopez-Poveda and P. T. Johannesen. Behavioral estimates of the contribution of inner and outer hair cell dysfunction to individualized audiometric loss. *JARO - Journal of the Association for Research in Otolaryngology*, 13(4):485–504, 2012.

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

T. V. Maele, S. Keshishzadeh, N. D. Poortere, I. Dhooge, H. Keppler, and S. Verhulst. The Variability in Potential Biomarkers for Cochlear Synaptopathy After Recreational Noise Exposure. *Journal of Speech, Language, and Hearing Research*, pages 1–18, 10 2021.

C. A. Makary, J. Shin, S. G. Kujawa, M. C. Liberman, and S. N. Merchant. Age-Related Primary Cochlear Neuronal Degeneration in Human Temporal Bones. *Journal of the Association for Research in Otolaryngology*, 12(6):711–717, 12 2011.

J. Märcher-Rørsted, G. Encina-Llamas, T. Dau, M. C. Liberman, P.-z. Wu, and J. Hjortkjær. Age-related reduction in frequency-following responses as a potential marker of cochlear neural degeneration. *Hearing Research*, 414:108411, 2 2022.

G. Mehraei, A. E. Hickox, H. M. Bharadwaj, H. Goldberg, S. Verhulst, M. C. Liberman, and B. G. Shinn-Cunningham. Auditory Brainstem Response Latency in Noise as a Marker of Cochlear Synaptopathy. *The Journal of Neuroscience*, 36(13):3755–3764, 3 2016.

B. C. Moore and B. R. Glasberg. Gap detection with sinusoids and noise in normal, impaired, and electrically stimulated ears. *Journal of the Acoustical Society of America*, 83(3):1093–1101, 1988.

B. C. Moore and B. R. Glasberg. A revision of Zwicker’s loudness model. *Acta Acustica united with Acustica*, 1996.

B. C. Moore, B. R. Glasberg, A. Varathanathan, and J. Schlittenlacher. A Loudness Model for Time-Varying Sounds Incorporating Binaural Inhibition. *Trends in Hearing*, 20:1–16, 2016.

B. C. J. Moore, R. W. Peters, and B. R. Glasberg. Detection of temporal gaps in sinusoids by elderly subjects with and without hearing loss. *The Journal of the Acoustical Society of America*, 92(4):1923–1932, 1992.

A. J. Oxenham. Predicting the Perceptual Consequences of Hidden Hearing Loss. *Trends in Hearing*, 20:1–6, 2016.

A. Parthasarathy and S. G. Kujawa. Synaptopathy in the Aging Cochlea: Characterizing Early-Neural Deficits in Auditory Temporal Envelope Processing. *The Journal of Neuroscience*, 38(32):7108–7119, 8 2018.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32(NeurIPS), 2019.

C. J. Plack, D. Barker, and G. Prendergast. Perceptual consequences of ”hidden” hearing loss. *Trends in Hearing*, 18:1–11, 2014.

G. Prendergast, H. Guest, K. J. Munro, K. Kluk, A. Léger, D. A. Hall, M. G. Heinz, and C. J. Plack. Effects of noise exposure on young adults with normal audiograms I: Electrophysiology. *Hearing Research*, 344:68–81, 2 2017.

A. F. Ryan and P. Dallos. Effect of absence of cochlear outer hair cells on behavioural auditory threshold. *Nature*, 253(5486):44–46, 1 1975.

M. R. Saddler, R. Gonzalez, and J. H. McDermott. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, 12(1), 2021.

R. H. Sanchez and T. Dau. Modeling spectro-temporal modulation perception in normal-hearing listeners. *Proceedings of the INTER-NOISE 2016 - 45th International Congress and Exposition on Noise Control Engineering: Towards a Quieter Future*, (10):1729–1740, 2016.

G. H. Saunders and M. P. Haggard. The Clinical Assessment of Obscure Auditory Dysfunction— 1. Auditory and Psychological Factors. *Ear and Hearing*, 10(3):200–208, 6 1989.

R. Schaette and D. McAlpine. Tinnitus with a Normal Audiogram: Physiological Evidence for Hidden Hearing Loss and Computational Model. *Journal of Neuroscience*, 31(38):13452–13457, 9 2011.

B. A. Schneider and S. J. Hamstra. Gap detection thresholds as a function of tonal duration for younger and older listeners. *The Journal of the Acoustical Society of America*, 106(1):371–380, 7 1999.

B. A. Schneider, M. K. Pichora-Fuller, D. Kowalchuk, and M. Lamb. Gap detection and the precedence effect in young and old adults. *The Journal of the Acoustical Society of America*, 95(2):980–991, 2 1994.

M. Schönwiesner and R. J. Zatorre. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14611–14616, 2009.

H. F. Schuknecht and R. C. Woellner. An Experimental and Clinical Study of Deafness from Lesions of the Cochlear Nerve. *The Journal of Laryngology & Otology*, 69(2):75–97, 2 1955.

L. A. Shaheen, M. D. Valero, and M. C. Liberman. Towards a Diagnosis of Cochlear Neuropathy with Envelope Following Responses. *JARO - Journal of the Association for Research in Otolaryngology*, 16(6):727–745, 2015.

M. J. Shailer and B. C. Moore. Gap detection as a function of frequency, bandwidth, and level. *Journal of the Acoustical Society of America*, 74 (2):467–473, 1983.

K. B. Snell. Age-related changes in temporal gap detection. *The Journal of the Acoustical Society of America*, 101(4):2214–2220, 4 1997.

H. Spoendlin and A. Schrott. Analysis of the human auditory nerve. *Hearing Research*, 43(1):25–38, 1989.

C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11141 LNCS:270–279, 2018.

S. E. Trehub, B. A. Schneider, and J. L. Henderson. Gap detection in infants, children, and adults. *The Journal of the Acoustical Society of America*, 98(5):2532–2541, 11 1995.

K. L. Tremblay, A. Pinto, M. E. Fischer, B. E. K. Klein, R. Klein, S. Levy, T. S. Tweed, and K. J. Cruickshanks. Self-Reported Hearing Difficulties Among Adults With Normal Audiograms. *Ear & Hearing*, 36(6):e290–e299, 11 2015.

M. Valero, J. Burton, S. Hauser, T. Hackett, R. Ramachandran, and M. Liberman. Noise-induced cochlear synaptopathy in rhesus monkeys (Macaca mulatta). *Hearing Research*, 353:213–223, 9 2017.

L. M. Viana, J. T. O'Malley, B. J. Burgess, D. D. Jones, C. A. Oliveira, F. Santos, S. N. Merchant, L. D. Liberman, and M. C. Liberman. Cochlear neuropathy in human presbycusis: Confocal analysis of hidden hearing loss in post-mortem tissue. *Hearing Research*, 327: 78–88, 9 2015.

J. P. Walton. Timing is everything: Temporal processing deficits in the aged auditory brainstem. *Hearing Research*, 264(1-2):63–69, 2010.

P. Wu, L. D. Liberman, K. Bennett, V. de Gruttola, J. T. O'Malley, and M. C. Liberman. Primary Neural Degeneration in the Human Cochlea: Evidence for Hidden Hearing Loss in the Aging Ear. *Neuroscience*, 407:8–20, 5 2019.

P.-z. Wu, J. T. O'Malley, V. de Gruttola, and M. C. Liberman. Age-Related Hearing Loss Is Dominated by Damage to Inner Ear Sensory Cells, Not the Cellular Battery That Powers Them. *The Journal of Neuroscience*, 40(33):6357–6366, 8 2020.

P.-z. Wu, J. T. O'Malley, V. de Gruttola, and M. C. Liberman. Primary Neural Degeneration in Noise-Exposed Human Cochleas: Correlations with Outer Hair Cell Loss and Word-Discrimination Scores. *The Journal of Neuroscience*, 41(20):4439–4447, 5 2021.

M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton. On rectified linear units for speech processing. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 3517–3521, 2013.

F. G. Zeng, S. Oba, S. Garde, Y. Sininger, and A. Starr. Temporal and speech processing deficits in auditory neuropathy. *NeuroReport*, 10 (16):3429–3435, 1999.

F. G. Zeng, Y. Y. Kong, H. J. Michalewski, and A. Starr. Perceptual consequences of disrupted auditory nerve activity. *Journal of Neurophysiology*, 93(6):3050–3063, 6 2005.

M. S. A. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *The Journal of the Acoustical Society of America*, 126(5):2390–2412, 11 2009.

Appendix A Hyper Parameters

The DNN model architecture used for the results in the main paper are a product of a series of exploratory experiments. These experiments are described in this appendix along with other additional considerations not included in section 2.3.

A.1 Model size

A common saying in general modelling, is the simpler the better. The models used by the papers performing similar studies to the present study used relatively large models (Saddler et al., 2021; Francl and McDermott, 2022). A smaller model is more efficient and faster to both train and evaluate, however reducing the size of a model might reduce it's ability to learn complex properties of the given data. Therefore, the following experiments was conducted. 1. We varied the shape of the kernels to from quadratic, to be longer in the temporal domain and in size generally. 2. we evaluated the effect of the number of channels, i.e. number of kernels per convolution block. 3. we halved the number of linear neurons in the hidden linear layer.

All the model comparisons was based on the simulated gap threshold for normal hearing, as well as the sensitivity to CS, in terms of resemblance with the NH and NP groups from (Zeng et al., 2005).

		Block 1	Block 2	Block 3	Block 4	Block 5
L3	Kernel	[1, 16]	[1, 12]	[1, 8]	[1, 4]	[1, 4]
	Stride	[1, 3]	[1, 2]	[1, 1]	[1, 1]	[1, 1]
	P.Kernel	[1, 3]	[1, 3]	[1, 3]	[1, 3]	[1, 3]
	P.Stride	[2, 2]	[2, 2]	[2, 2]	[2, 2]	[2, 2]
L4	Kernel	[5, 5]	[3, 3]	[3, 3]	[3, 3]	[3, 3]
	Stride	[2, 2]	[1, 1]	[1, 1]	[1, 1]	[1, 1]
	P.Kernel	[3, 3]	[3, 3]	[3, 3]	[3, 3]	[3, 3]
	P.Stride	[2, 2]	[2, 2]	[2, 2]	[2, 2]	[2, 2]
L5	Kernel	[1, 5]	[1, 3]	[1, 3]	[1, 3]	[1, 3]
	Stride	[1, 2]	[1, 1]	[1, 1]	[1, 1]	[1, 1]
	P.Kernel	[1, 3]	[1, 3]	[1, 3]	[1, 3]	[1, 3]
	P.Stride	[2, 2]	[2, 2]	[2, 2]	[2, 2]	[2, 2]
L6	Kernel	[5, 16]	[3, 12]	[3, 8]	[3, 4]	[3,4]
	Stride	[2, 3]	[1, 2]	[1, 1]	[1, 1]	[1, 1]
	P.Kernel	[3, 3]	[3, 3]	[3, 3]	[3, 3]	[3, 3]
	P.Stride	[2, 2]	[2, 2]	[2, 2]	[2, 2]	[2, 2]

TABLE 1
The kernel size and stride for each evaluated model in section A.1.1. P.Kernel and P.Stride is short for pooling kernel size and stride, respectively.

A.1.1 Kernel shape

We tested 4 kernel shapes with fixed but different strides. The sets of layer parameters is seen in table 1. We tested 4 training initialisations and evaluated the models by their mean thresholds, their inter-model variance, resemble to human NH GDTs and their sensitivity to CS for each set of model parameters.

The initial model tested, was "L4", which uses the same kernel sizes and strides as used in (Haro et al., 2020). However, this model quickly obtained better-than-human thresholds in the simulated NH condition, while it little-to-none increase in threshold for the CS conditions. This led to removing the spectral dimension of the kernels, i.e. use model "L5". However, this model was unable to obtain close-to-human performance, and showed large inter-model variance for the CS conditions. This led to expanding the temporal filters in size, i.e. model "L3". This yielded good results and was kept as preferred model structure. To test whether this was an effect of increased temporal kernel length, we also tested "L6", which is a combination of "L3" with the spectral sizes from "L4".

The models with spectral kernels was much less sensitive to CS than the models using temporal-only kernels. To exemplify, the L4 model quickly obtained better-than-human performance in terms of GDT for the NH condition.

Based on these findings, the "L3" model parameters was elected throughout the study.

A.1.2 Channel factor

We evaluated channel factors 4, 8, 16 and 32 in terms of the minimum GDT across presentation level. The GDT for each presentation level in the range of 20 to 80 dB SPL is shown for each channel factor and grade of CS in fig. 9. The annotations mark the minimum GDT of the combined model ensemble for each model and impairment condition, respectively. We can visually see that the models with $C = 16$ perform equally good as the $C = 8$ models for the NH condition, but are more sensitive to CS. Our goal is to select a model that performs as good as the NH human group, but it also sensitive to CS. A third measure taken into account was the resemblance with the human data. From the current comparison, a channel factor of $C = 32$ was most alike the human data in terms of stable change of GDT across level. The minimum GDT for the NH condition and CS seemed reasonable within what was achieved by the other channel factors.

A.1.3 Size of hidden layer

We tested a model further reduced by size with half the neurons (256) in the fully connected layer following the convolution blocks. However, the model showed larger fluctuation across level and slightly higher GDTs compared to the larger model with 512 neurons. Based on this subtle comparison, we chose the larger model.

A.2 Training data

We considered the effect of the generated datasets by training models on the BBN training data and all of the babble-noise conditions, respectively. To analyse how each dataset generalised to the BBN task, we evaluated the training loss along with the validation loss, which we obtained by testing the model on a separate BBN test set. This validation set was not part of the data sets used for obtaining the gap thresholds shown in section 3, but was similar with gap position in the centre of the carrier. All datasets was evaluated by training 4 models for each, with the same hyper-parameters as used in the main study.

Checking the training loss and validation loss for each of the individual training sets, we didn't see a big change in training loss decay rate among the babble-noise training datasets. Neither did the validation loss substantially differ among those 6 datasets, which after 200 training iterations all reached around 3 in value. Most notably was the 5-talker babble-noise data set, were the validation loss increased slower than the other data set, until about 50 training iterations. However, after 50 training iterations, the validation loss was similar to the other n-talker babble noise sets. The training loss was similar for all of the babble noise datasets.

A different result was obtained from the BBN dataset. For the first 10 training iterations, the validation loss was below the training loss, indicating the test set to be ”easier” for the model while training. After 10 training iterations, it started to increase, following a log-like function. The validation loss after 200 training iterations averaged to 0.5, thus substantially better than for any of the babble-noise data sets alone.

In overall conclusion, using a subset of the complete training dataset, the DNN model seem to overfit after about 50 training iterations. It should however be considered that the evaluated subsets are vastly smaller in size, compared to the full dataset.

A.3 Training iterations

Leading from the above experiment, we wanted to observed the behaviour of the model in terms of GDT and sentivity to CS as a function of training iteration.

To see whether an effect could be observed during training, we saved a checkpoint for every 20 training iterations over the course of 400 complete iterations. Then, the GDT for the different degrees of CS with NH audiogram was obtained for the DNN model weights at each checkpoint. This was done for 4 individual model initialisations at a presentation level of 40 dB SPL. The average GDT is depicted in the upper panel of fig. 10. We also wanted to test the change in GDT due to CS, i.e. the sensitivity to CS as a function of training iteration. This is shown in the lower panel of fig. 10. The difference was computed simply by

$$\Delta \mathrm{GDT}(c) = \mathrm{GDT}(c) - \mathrm{GDT}(\mathrm{NH}) \,, \tag{19}$$

where c is the given simulation condition.

We observed that the lowest GDT for the NH condition without CS was for the checkpoint at 40 data iterations. Furthermore, this showed the locally highest sensitivity to CS. Based on this result, we chose this training length as the optimal selection, thus the DNN model weights after 40 data iterations was used in section 3.

A.4 Effect of fibre type discrimination

We also tested the effect of having each fibre type as a designated input channel, (see eq. (1)). However, in the implementation of the convolutional layers, we didn't use depth-wise convolution, thus all inputs was convolved then summed for each output channel without training kernels specific for one channel or with weights specifically for the combination of the grouped kernels. The DNN model architecture is only changed in the first convolutional layer, where two channels, i.e. two additional sets of kernels are added. When comparing the model performance, both in terms of the rate of decay of the training loss and the obtained GDTs for both NH and the simulated grades of CS we found no substantial difference. We therefore decided to not investigate this parameter further. It's an open question on whether the AN-fibres are routed to fibre-type specific processing centres of the brain, and the role of the different fibre types is not completely understood. The PyTorch library includes the possibility of handling ”groups”, in their implementation of convolutional layers, which is equivalent to designing parallel layers side-by-side for each group. By using n groups, the only constraint is that the number of kernels must be a multiple of n.

A.5 Threshold computation

We evaluated three methods for inferring a gap threshold for the trained DNN model (see section 2.2). All involved fitting a psychometric function to either the direct output of the model, or the percentage of correct hits with a fixed or ideal decision criterion.

Using the model-output method generally resulted in higher GDTs. Percentage-correct with the fixed threshold at 0.5 generally lowered the threshold, while percentage-correct with an estimated ideal decision criterion (eq. (5)) resulted in the lowest threshold. It seems unclear to which of the methods are most naturalistic, and given that we don't no the correlation between the DNN model activations and the actually ”implementation” in the auditory brain stem, the basis for comparison is vague. However, the model-output method showed a curvature for the NH case that resembled the human results from (Zeng et al., 2005) better. However, when the mean DNN model output doesn't change substantially across gap length, the fitting of the psychometric function breaks and thus for lower levels, we found GDTs jumping from 2 to 35 ms for presentation levels at 20 dB SPL and 25 dB SPL. A similar behaviour was observed for the percentage correct method using a fixed decision criterion. As described earlier, the DNN models obtained a bias towards ”gap” for the lower presentation levels, thus the model would get 100 % hits for all samples with a gap at that level, but also 100 % false alarms for the no-gap samples.

The first two methods apply no knowledge of the models response to the other samples, which resembles a yes/no procedure, where the answer of the subject is based on one view of the stimulus, with no direct base of comparison. The third method is more comparable to a n-AFC procedure, as the DNN model output is used as the basis for setting an optimal decision criterion. During an n-AFC, the subject is usually presented with gap lengths that are much longer than the expected threshold. This enables the subject to learn the internal response given a gap and no-gap.

Based on the above findings, we used the intersection of the psychometric functions fitted to percentage correct using an estimated ideal decision criterion throughout the results presented in the main paper.

Appendix B
Gap Detection

The following section acts as a micro review on gap detection, carried out as part of the present study. Gap detection has generally been used as a mean of investigating the temporal acuity of the auditory system in various species. Various potential conditions of the auditory periphery that might affect gap detection thresholds has been evaluated. This includes effects of ageing (Schneider et al., 1994; Schneider and Hamstra, 1999; Trehub et al., 1995; Snell, 1997; He et al., 1999; Walton, 2010) and hearing impairment (Fitzgibbons and Wightman, 1982; Irwin et al., 1981; Moore and Glasberg, 1988; Moore et al., 1992; Zeng et al., 2005). Gap detection has furthermore been carried out with different carrier characteristics.

In general the stimuli used in the gap detection studies included in this study were varied along four dimensions; 1. the **intensity** of the markers, 2. the **duration** of the markers, 3. **spectral width** of the

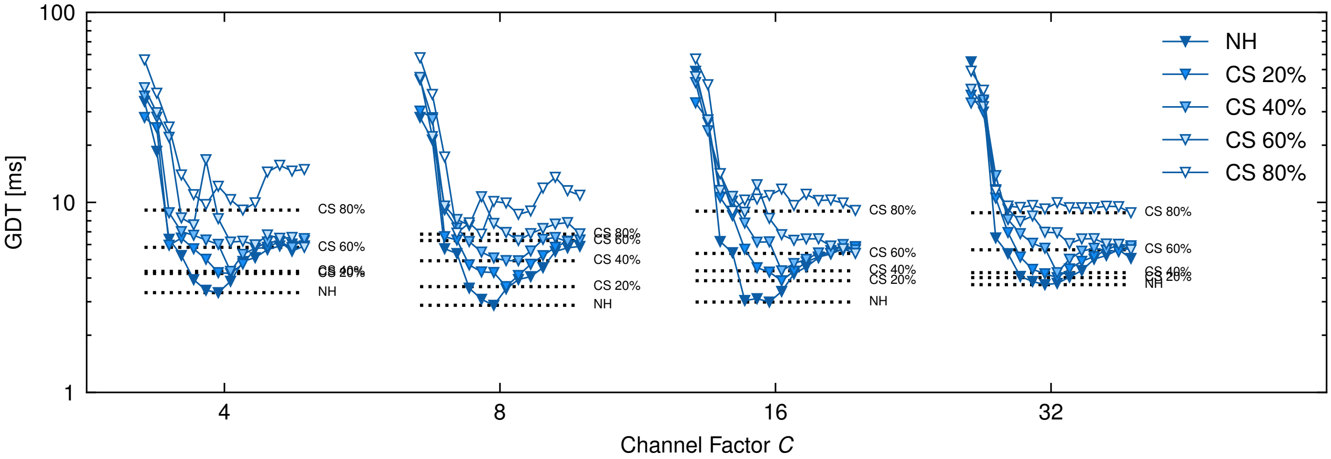

Fig. 9. Each GDT curve depicts the mean of 4 individually trained models for each of the four values of C. The curves are plotted from 20 to $80\,\mathrm{dB\,SPL}$ in steps of $5\,\mathrm{dB\,SPL}$ and for each grade of simulated CS, illustrated by the colour of the filling in the markers. The dotted annotations depict the minimum GDT of each ensemble for each hearing impairment condition.

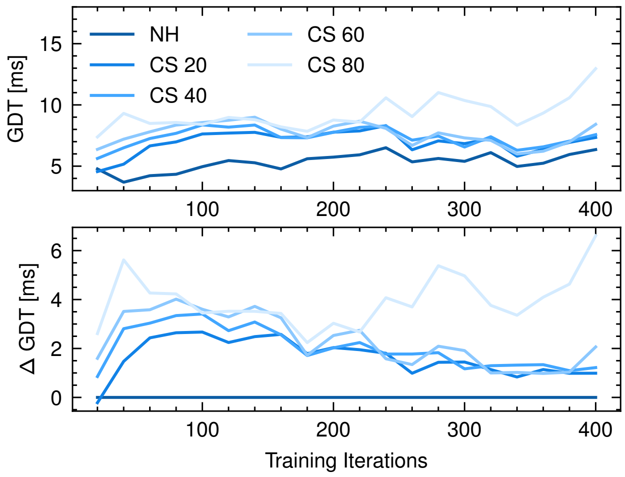

Fig. 10. Obtained GDT (upper panel) and GDT-difference for each training checkpoint averaged for 4 DNN models presented at $40\,\mathrm{dB\,SPL}$. The GDT-difference is condition coloured as the above panel subtracted by the simulated NH GDT.

stimulus, ranging from pure-tones to broadband noise, 4. frequency range or **centre frequency**.

Usually, the gap detection thresholds are evaluated over one or more experiment parameters, specific for the given study. For instance, studies have estimated the effect of varying the centre frequency of the carrier (Fitzgibbons and Wightman, 1982; Formby and Forrest, 1991; Moore et al., 1992; Florentine et al., 1999), the carrier intensity (Irwin et al., 1981; Schneider et al., 1994; Zeng et al., 1999, 2005) and carrier duration (Schneider and Hamstra, 1999).

We decided for the present study to focus on simulating the gap detection setup used in Zeng et al. (2005), based on the simple stimuli and large difference between the groups and with the NP having the largest probability of being affected by the CS.

Appendix C Babble Noise

For the training a data, a collection of babble noise was created. Synthetic babble noise already exists within the ICRA noise dataset (Dreschler et al., 2001). However, the noise is modulated as a 3-band vocoder and thus it doesn't sound that natural. Therefore, the decision was made to create a babble noise data set, using the HCHR Map Task Corpus (Anderson et al., 1991) for generating babble noise of various count of speakers. The corpus contains two-person dialogues of one person primarily speaking at the time (their might be brief periods where the two speakers overlap). In the recordings, two participants takes turn on calmly describing some path along a map.

Six samples of babble noise was created for 3, 4, 5, 6, 7 and $8\,\mathrm{speakers}$, respectively. Thus, as an example, the 3-speaker babble noise contained 3-different recordings of the HCHR Map Task corpus. The recordings were added together raw, meaning that no additional processing was applied. The recordings from the corpus were of different length, thus the recordings was truncated by the shortest one included, ensuring that one of the speech recordings wouldn't include a lesser count of speakers towards the end of the mixed babble noise.

The speech sequences used for mixing the babble noise was randomly selected among the available recordings in the source corpus. The final mixed babble noises had a duration of between 3- and 4 minutes long, respectively.

C.1 Gap detection preparation

The babble noise dataset was used for training the deep learning models, simulating the internal decision variable following the auditory nerve, used for determining whether a gap was present or not. The babble noise clips of length N was sampled randomly (using a uniform distribution over the range 0 to $N - K$, with K being the number of samples in the sub-sampled clip. For each sample, a gap and no-gap version where created.

The gap condition was created by sampling gap length and gap position uniformly across samples. The gap position was sampled from $100\,\mathrm{ms}$ after stimuli onset to $100\,\mathrm{ms}$ before the offset. Similarly was the gap length sampled in the range 0.5 to $100\,\mathrm{ms}$. For each number of speakers in the babble, 400 samples was cut, resulting in 800 stimuli. With 6 speaker conditions, this resulted in 2400 samples with gaps and 2400 samples without. Creating a total of 4800 unique sound files. To visually inspect the range of gap position and duration, see fig. 11.

The sampled gap durations and positions are indeed uniformly distributed within the defined range, fig. 12 shows histograms of all durations and positions sampled for all n-talker conditions, respectively.

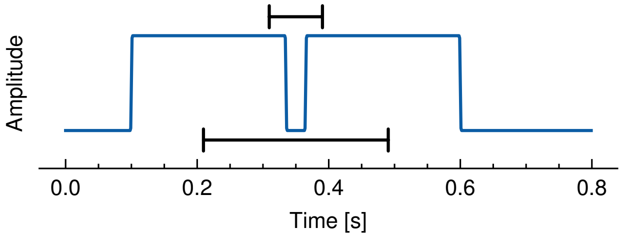

Fig. 11. The envelope for a gap condition, with position perfectly centred at $350\,\mathrm{ms}$. The two black lines indicate the range of gap centre placement (the line below) and gap duration (the line above).

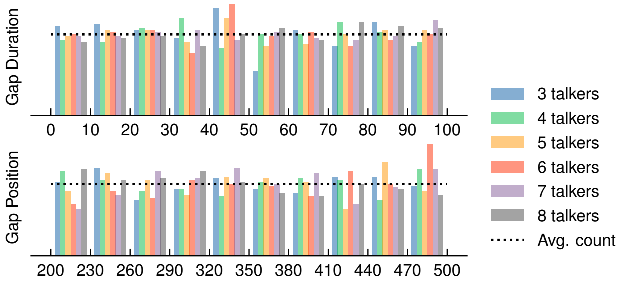

Fig. 12. Distributions of gap length and duration in the babble noise dataset, depicted for n-talker condition. The average count (dotted lines) would be the perfect distribution across condition, with 30 samples in each bin, as the total count of samples with gaps for each condition is 300.