From Spikes to Speech: NeuroVoc
A Biologically Plausible Vocoder Framework for Auditory Perception and Cochlear Implant Simulation

Jacob de Nobel, Jeroen J. Briaire[b], Thomas H.W. Bäck[a], Anna V. Kononova[a], Johan H.M. Frijns[b,c,d]

[a]*Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, Leiden, Netherlands*
[b]*Department of Otorhinolaryngology, Leiden University Medical Center, Albinusdreef 2, Leiden, Netherlands*
[c]*Leiden Institute for Brain and Cognition, Wassenaarseweg 52, Leiden, Netherlands*
[d]*Bioelectronics group, EEMCS, Delft University of Technology, Mekelweg 4, Delft, Netherlands*

Keywords: Hearing Loss, Cochlear Implants, Auditory Periphery, Neural Model, Vocoder, Auditory Perception, Auditory Nerve, Neural Decoding, Signal Processing, Phenomological Model

1. Introduction

In Cochlear Implant (CI) research, vocoders are often used as simulators to mimic how sound is processed and heard through a CI (Cychosz et al., 2024). Traditionally, the vocoder, which is a conjugation of the words *voice* and *encoder*, is a signal processing method used to break down and reconstruct speech material for efficient telecommunication (Dudley, 1939). Specifically, this so-called channel vocoder works by extracting the temporal envelope of an audio signal for a limited set of frequency bands. Transmitting these envelopes only requires several samples per second, whereas the original audio requires thousands of samples per second. On the receiving end, these envelopes, combined with a specific carrier signal, can be reconstructed into intelligible speech. Because the vocoder can be precisely controlled and parameterized, it is a powerful tool for studying how sound is perceived. This is crucial for understanding how cochlear implants transform acoustic signals and how listeners, especially individuals with normal hearing (NH) in studies, might perceive speech or other sounds as if they were CI users

(Shannon et al., 1995).

Cochlear implants are medical devices that restore hearing to individuals with severe to profound deafness and are considered the most successful neuroprosthetic device developed to date (Kansaku, 2021), with over one million people having been implanted worldwide (National Institute on Deafness and Other Communication Disorders, 2024). One of the primary goals of cochlear implants is to partially restore access to speech information, thereby enabling effective communication. While successful, outcomes vary significantly from patient to patient due to factors such as age and the duration, cause, and type of hearing loss. Consequently, response data for studies with CI-users often experience high subject-level variability, which is hard to contain or isolate (Blamey et al., 2012). Furthermore, clinical trials can only rely on a relatively small population of CI users. This provides a challenging environment in which to evaluate the plethora of design choices available for developing a CI (Cychosz et al., 2024).

As mentioned, an alternative approach is to present Normal Hearing (NH) listeners with signals processed by a channel vocoder to emulate the signal as if perceived by CI users. In practice, the listener can then attempt to recover the content of the original signal, for example, in comprehension tasks (Shannon et al., 1995). This general framework has become essential in studying hearing loss and allows us to better understand how individuals with CI perform auditory, speech, and language tasks. Moreover, in addition to providing a much larger patient population for conducting trials, it allows for testing specific experimental conditions in isolation (Cychosz et al., 2024). This has enabled researchers to conduct several studies that would have been challenging or impossible to conduct without relying solely on CI users. For example, studies using vocoders have been employed to reveal how degraded speech affects language development (Newman et al., 2020) and how deeper cochlear implant (CI) insertion depth enhances speech perception (Rosen et al., 1999; Shannon et al., 1998). Additionally, it provides a way for NH individuals to experience some aspects of the sound quality of Electrical Hearing (EH). However, it should be noted that a vocoder does *not* simulate the experience of wearing a CI. Aside from the social and practical implications of being implanted, there are inherent differences between the healthy acoustic hearing system and EH that the signal processing strategies of a vocoder cannot capture (Cychosz et al., 2024).

In a parallel branch of CI research, the CI listening experience is studied from a different perspective: simulation with computer models (Rattay, 1986; Frijns et al., 1995). This takes another approach to avoid conducting physical experiments with CI users, which requires considerable effort from the human subjects involved. In addition, digital twins have allowed researchers to model specific effects of the auditory system in response to (electrical) stimulation. For example, 3D models helped uncover the current spread throughout the cochlea when stimulated with a given electrode array (Kalkman et al., 2022). Moreover, model studies can provide insight into the human hearing system at the single-fiber level (Bruce et al., 1999; Rattay et al., 2001). While inherently an abstraction, models can provide a powerful way to study specific effects, enabling a depth and scale of investigation that is often not possible with animal models—and especially not with live human subjects (Hanekom and Hanekom, 2016).

As previously mentioned, vocoders do not capture all the important biophysical aspects related to perception in CI users, as they are based solely on signal processing techniques. Standard channel vocoders do not consider effects such as single fiber refractoriness, electrode interaction, and electrode-to-neural interface. In addition, vocoder design is often specific to a given implant or speech coding strategy, making evaluating a strategy change problematic (Cychosz et al., 2024). El Boghdady et al. (2016) proposed a hybrid between the modelling and vocoder-centric approach. This work used a simple population-based Auditory Nerve Fiber (ANF) model as a preprocessing step to the standard vocoder used by the Advanced Combinatorial Encoder (ACE) strategy. This makes it possible to study the effects of newly developed coding strategies within the same framework as already established methods. While El Boghdady et al. (2016) uses a neural model only as a preprocessing step to a standard channel vocoder, the general methodology follows that of *neural decoding* (Johnson, 2000). This approach is analogous to those of Pasley et al. (2012); Akbari et al. (2019), which utilize ECoG (Penfield and Jasper, 1954) recordings to reconstruct intelligible speech. Other approaches (Park et al., 2023; Daly, 2023) have used fMRI readings to reconstruct complex musical pieces from brain signals using deep learning techniques.

Besides the fact that no live patients are required to conduct experiments

in *model-based* neural decoding, an additional advantage is that there is no need to rely on an imperfect signal, such as those collected via fMRI or ECoG. Simulation with models yields exact spike timings per fiber, allowing for precise measurement of the information transferred to the auditory nerve (Johannesen et al., 2022). The recent work by Leclère et al. (2023) proposed an information-theorectic framework to asses the information contained in the simulated spiking response of a computational model of the implanted auditory nerve. Their model started from the electrode-neural interface (ENI), i.e., from an electrodogram. It then used optimal reconstruction filters to reconstruct the temporal envelope of amplitude and rate-modulated reference signals from the simulated spike trains, based on the approach by Warland et al. (1997).

In this work, we propose a *general methodology* for decoding the output of neural models into sound. In this sense, we can leverage the advancements of contemporary models (Bruce et al., 2018; Kalkman et al., 2022; Lyon et al., 2011; de Nobel et al., 2024) to develop a biophysically plausible vocoder that reconstructs sound from neurograms, time-frequency representations of auditory nerve activity (Hines and Harte, 2012). Leveraging the relationship between the neurogram and the spectrogram, our method employs an inverse Fourier transformation for reconstruction. This, in principle, allows for *any* neurogram-generating source to be used in the simulation process. We demonstrate this using two different ANF models for normal and electrical hearing, without requiring any ad-hoc parameter tuning. This flexibility also allows for the variation of any parameters in these models to match specific experimental conditions, enabling the evaluation of, for example, new speech coding strategies or implant designs within the same computational framework. Since the input signals are in the same domain as the reconstructed signals, i.e., sound, information-theoretic approaches can be applied to quantify the effects of such developments numerically, which is vital for automated development (Bäck et al., 2023). Our main contributions are:

- We propose a flexible vocoder framework that reconstructs sound from simulated auditory nerve activity using classic signal processing.
- The framework supports interchangeable auditory models, enabling direct comparison between normal hearing and cochlear implant conditions without requiring model-specific vocoders.

- We demonstrate that the vocoder captures characteristic differences between models.
- We evaluate perceptual intelligibility using an online Digits-in-Noise (DIN) test and show that our results align with clinical benchmarks.

The structure of this paper is as follows. Section 2 introduces the necessary preliminaries, including relevant background and model components. Section 3 provides a detailed outline of the proposed framework. Section 4 presents two experiments that evaluate the reconstructed sound, including a perceptual assessment using the Digits-in-Noise (DIN) test. Finally, Section 6 concludes the paper by discussing the findings and implications.

2. Preliminaries

2.1. Short-Time Fourier Transform

Let $x[t]$ denote a discrete-time signal of length T, sampled at a rate of f_s samples per second, where $t = 0, 1, \ldots, T-1$. The Short-Time Fourier Transform (STFT) provides a time-frequency representation of $x[t]$ by analyzing short overlapping segments of the signal, allowing the frequency content to be tracked over time (Oppenheim, 1999). The STFT is computed by multiplying $x[t]$ with a window function $w[t - t_k]$ centered at time frame t_k, treating $x[t] = 0$ for t outside $[0, T - 1]$, followed by a Fourier transform. Formally, the STFT and its inverse (ISTFT) are defined as:

$$\mathrm{STFT}(x[t]) = X[t_k, f_i] = \sum_{t=0}^{T-1} x[t] w[t - t_k] e^{-j2\pi \frac{f_i}{f_s} t}, \tag{1}$$

$$\mathrm{ISTFT}(X) = \hat{x}[t] \qquad = \frac{\sum_k \hat{x}_k[t - t_k] w[t - t_k]}{\sum_k w^2[t - t_k]}, \tag{2}$$

where f_i represents the physical frequency in Hertz (Hz), f_s is the sampling rate, and $\hat{x}_k[t]$ is the inverse Fourier transform of $X[t_k, f_i]$. When appropriately overlapping windows are used (e.g., satisfying the constant-overlap-add condition (Allen, 1977)), the ISTFT allows for perfect reconstruction of the original signal $x[t]$.

The spectrogram, or more specifically, the *magnitude spectrogram* $S[t_k, f_i]$, is given by the magnitude $|X[t_k, f_i]|$ and represents the amplitude spectral

density of the signal. This provides a compact representation that can be used for visualization, feature extraction, and further analysis.

2.2. Griffin-Lim Phase Reconstruction

While the complex-valued STFT $X[t_k, f_i]$ is invertible, the magnitude spectrogram $S[t_k, f_i]$ alone discards phase information, making direct inversion impossible. The Griffin-Lim algorithm (Griffin and Lim, 1984) provides an iterative procedure to estimate the missing phase. Given $S[t_k, f_i]$, Griffin-Lim iteratively refines a complex STFT $\hat{X}[t_k, f_i]$ such that:

$$|\hat{X}[t_k, f_i]| \approx S[t_k, f_i]$$

and the inverse STFT of $\hat{X}$ corresponds to the STFT of a valid time-domain signal, i.e.:

$$\hat{X} \approx \text{STFT}(\text{ISTFT}(\hat{X})).$$

Once the algorithm converges to a stable solution, or after a predefined number of iterations, the estimated $\hat{X}$ can be used to reconstruct $\hat{x}[t]$. This works because the overlapping windows in the STFT introduce redundancy, causing time-frequency components to share information. In particular, adjacent frames partially overlap in time, and spectral leakage spreads energy across nearby frequencies, allowing the Griffin-Lim algorithm to iteratively estimate phase from the shared structure in the magnitude spectrogram.

2.3. Mel Scale

The Mel scale, named after the word *melody*, is a perceptual scale of equally spaced pitch intervals (Stevens et al., 1937). It reflects the nonlinear sensitivity of the human auditory system and is typically defined as a quasi-logarithmic function of acoustic frequency. The scale is constructed such that equal distances on the Mel axis correspond to perceptually uniform pitch intervals across a specified frequency range. Since it is derived from perceptual experiments in healthy hearing individuals, several implementations exist; in this work, we use the widely adopted version from Slaney (1998).

A Mel spectrogram differs from a standard spectrogram in how it represents frequency. Rather than linearly spaced frequency bins, the frequency axis is distributed according to the Mel scale. A linear-frequency spectrogram can be converted into a Mel spectrogram by applying a Mel filterbank, which

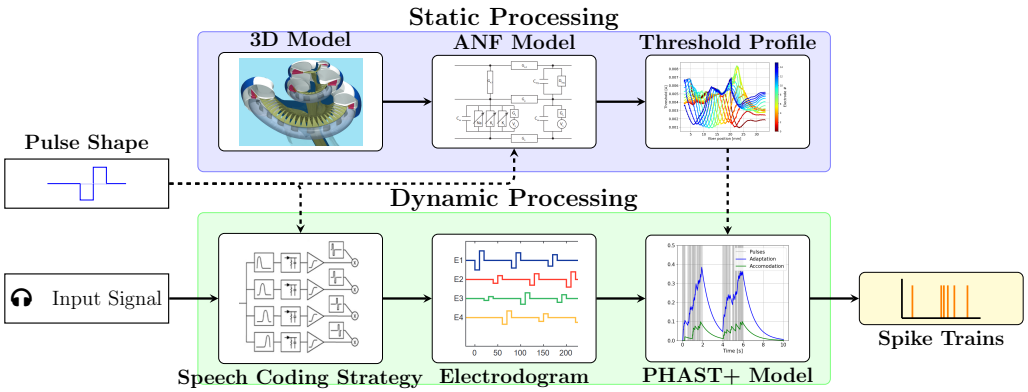

Figure 1: Diagram of the EH modelling pipeline, consisting of a static and a dynamic part. Solid lines represent data flow, dashed lines denote the exchange of fixed information. The static part of the pipeline calculates a threshold profile for a specific 3D configuration of an implanted cochlea when stimulated with a predefined pulse shape. The dynamic part of the pipeline simulates a temporal response to an incoming input signal, producing spike trains.

projects the spectral content into the Mel frequency domain. This results in finer resolution at lower frequencies and coarser resolution at higher frequencies, yielding a more compact and perceptually meaningful representation, particularly effective in audio and speech processing.

2.4. Modelling Electrical Hearing

To model EH, we employ a modeling pipeline based on the work of Kalkman et al. (2022) and de Nobel et al. (2024), utilizing a cascade of biophysical and phenomenological models to generate spike trains for a simulated cochlear implant user. This modeling approach is illustrated schematically in Figure 1, consisting of two main components.

Static Processing. The first part of the pipeline, indicated with purple in the diagram, models the Electrode-Neuron Interface for a human-implanted cochlea under stimulation with a predefined stimulus waveform (pulse shape). This includes a 3D volume conduction model, which employs a boundary element method to simulate electrical fields in cochleae with arbitrary geometries implanted with multi-channel electrode arrays (Briaire and Frijns, 2000). The diagram in Figure 1 shows an example of such a geometry. This is followed by a deterministic Auditory Nerve Fiber model (Dekker et al., 2014; Kalkman et al., 2022). This simulates a non-linear double cable model

of a human bipolar High Spontaneous Rate fiber (Frijns et al., 2000; Briaire and Frijns, 2005) with Schwarz-Reid-Bostock kinetics (Schwarz et al., 1995). This is used to calculate the activation threshold of a fiber when stimulated by a given electrode contact with the predefined pulse shape. Applied for a set of n_f fibers and n_e electrode contacts, this produces a threshold profile, which is a $(n_f \times n_e)$ matrix of activation thresholds.

Dynamic Processing. Where the static part of the pipeline models a fixed stimulation threshold for a single stimulus waveform, the dynamic part simulates a complete temporal response to an incoming audio signal. This includes a Speech Coding Strategy (SCS), configured with the same number of electrode contacts as were used for modeling the 3D geometry, which generates an electrogram by processing the input signal. The electrodogram, also known as a pulse train, is a multivariate time series comprising n_e channels, where each channel represents the current level of an electrode contact at a specific point in time. This is then used as the input for the PHAST+ model (de Nobel et al., 2024), a computationally efficient version of the phenomenological model introduced in (van Gendt et al., 2016). This model converts the pulse trains as generated by an SCS into a simulated spike train, adding temporal behaviour on top of the deterministic thresholds calculated by the ANF model from the static part of the pipeline. The PHAST+ model uses these thresholds to determine the spiking behaviour of an ANF by incorporating the following temporal effects:

- **Accommodation:** A gradual increase in threshold due to sustained stimulation (Hodgkin and Huxley, 1952), modelled by a leaky integrator.
- **Adaptation:** A decrease in firing rate over time in response to prior spiking activity (Litvak et al., 2001), modelled as an increase in the threshold by a leaky integrator.
- **Refractoriness:** Temporary inability, or reduced ability of an ANF to fire following a recent spike (Yeomans, 1979), modeled as an (potentially infinite) increase of the threshold.
- **Stochasticity:** A stochastic activation threshold (Verveen and Derksen, 1968). Modelled by a random normal variable with a standard deviation of 5% of the deterministic threshold, it is used to randomly lower or increase the threshold slightly for each stimulus presentation.

- **Spontaneous Firing:** Stimulation-independent spontaneous firing behaviour (Kiang et al., 1966), modelled by a Poisson process which randomly causes an ANF to produce spikes. This was not included in de Nobel et al. (2024) and is added to the model specifically for this work. This parameter is set to a constant 50 spikes per second for all modeled fibers.

The code for the PHAST+ model is available as an open-source Python package[1] and includes several pre-processed threshold profiles for different cochlear geometries and electrode arrays.

2.4.1. Speech Coding Strategy

The speech coding strategy is taken from the Advanced Bionics Generic-Python-Toolbox (jabeim, 2025), modified for interoperability with PHAST+. The code models the Spectral Resolution (SpecRes) strategy, which is a research version of the HiRes Fidelity 120 processing strategy (Nogueira et al., 2009). The strategy uses asynchronous sequential pulses like Continuous Interleaved Sampling (CIS) (Wilson et al., 1991) technique and works via the same fundamental principles:

- The incoming signal is divided into several frequency bands using a bandpass filterbank.
- Each band's envelope (the slow-changing amplitude of the signal) is extracted, discarding the fine structure (fast oscillations).
- These envelopes are then used to modulate a train of biphasic electrical pulses.
- The pulses are delivered sequentially across electrodes, one at a time, in rapid succession. This prevents overlapping stimulation and reduces channel interaction.

SpecRes is used to process incoming acoustic signals and generate pulse trains for 16 electrode contacts. Unlike CIS, which assigns one electrode per filter band, SpecRes utilizes current steering (Bonham and Litvak, 2008), which involves the pairwise stimulation of two adjacent electrodes. The strategy

[1] see: https://github.com/jacobdenobel/PHAST

uses Fast Fourier Transform (FFT)-based filtering to separate the incoming signal into 15 analysis bands. Due to the limited precision of these filter banks, the strategy effectively acts as a bandpass filter on the acoustic signal, limiting the frequency content from 306 Hz to 8 054 Hz. Analysis bands span two electrodes, and within each band, a spectral peak locator identifies dominant frequencies. This is used to determine a weighting scheme, where the electrode with its operating frequency closer to that of the estimated peak gets a larger weight. This enables the creation of so-called virtual electrode channels, which provide higher spectral resolution for the CI user (Bonham and Litvak, 2008). By default, SpecRes separates each analysis band into nine distinct steps between each electrode pair, resulting in a total of 120[2] unique virtual channels. For more details, we refer the interested reader to Nogueira et al. (2009).

2.5. DIN Test

The Digit-in-Noise (DIN) test (Smits et al., 2013) is a speech-in-noise hearing assessment that measures a listener's ability to recognize spoken digits (typically 0–9) presented against background noise. It is widely used for its simplicity, reliability, and suitability for remote or clinical settings. The test determines the signal-to-noise ratio (SNR) at which a person can correctly identify 50% of the digit triplets, providing an estimate of speech perception in noisy environments. Because it uses language-independent numerical stimuli, the DIN test is accessible across different populations (Polspoel, 2024) and has been shown to correlate well with traditional speech-in-noise tests (Kwak et al., 2021). It was originally developed as an online test, and Shehabi et al. (2025) found that the difference between clinical and online testing was not statistically significant for both Arabic and English language speakers.

3. Methods: The *NeuroVoc* Framework

Figure 2 presents the architecture of the proposed biologically inspired neural vocoder, *NeuroVoc*, which follows an encoder–decoder design. The encoder serves as a flexible simulation framework that can incorporate any neural population-based hearing model. The modular design enables substitution of the entire model as well as parametric manipulation, allowing the

[2] $15 \cdot 9 = 135$ containing 15 'duplicate' channels

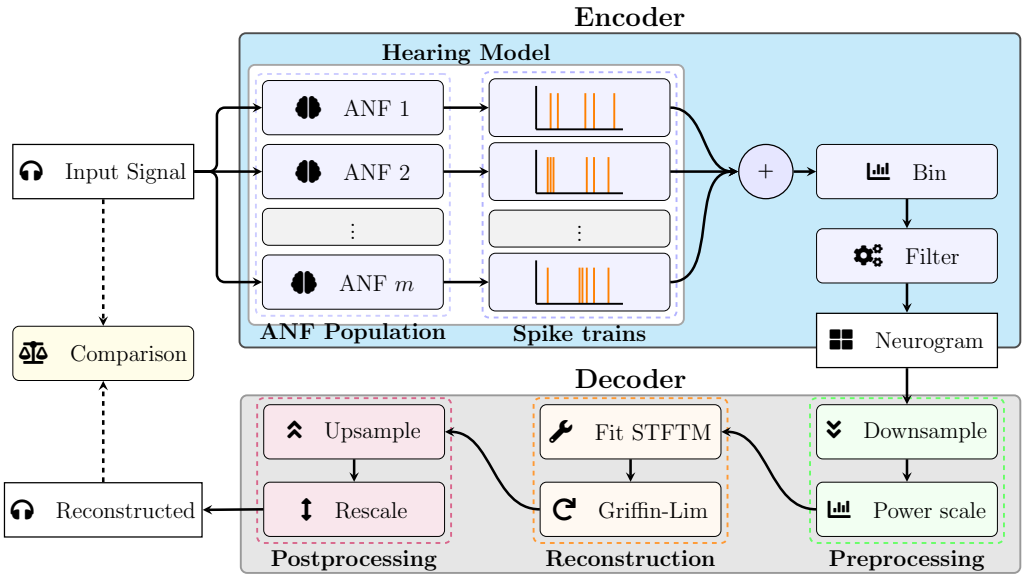

Figure 2: Schematic of the NeuroVoc architecture. An input sound is encoded by a population of variable auditory nerve fiber (ANF) models, producing spike trains that are binned and filtered to generate a neurogram. The decoder reconstructs the sound through spectral and temporal transformations, enabling comparison with the original signal.

simulation of diverse experimental conditions such as neural health, implant configuration, and coding strategy. The decoder reconstructs the acoustic signal using only the neurogram as input. The following sections provide a detailed description of each component.

Code availability. A Python implementation of the method presented in this work, complete with examples, along with all the code necessary to run the experiments and produce the figures included in this paper, is available open-source at https://github.com/jacobdenobel/NeuroVoc.

3.1. Encoder: Generating Neural Responses to Sound

The encoder simulates peripheral auditory processing using a population of auditory nerve fiber (ANF) models, each characterized by specific parameters, such as characteristic frequency (CF), spontaneous rate, and temporal response profile. These models receive the acoustic input and produce discrete spike trains that reflect the stimulus-driven firing behavior of individual nerve fibers. For this purpose, any model that simulates the peripheral auditory process can be used, as depicted by the white shaded area in Figure

2. Here, both the acoustic model of Bruce et al. (2018) (see Section 3.3) and the EH model presented in de Nobel et al. (2024) (see Section 3.4) are used to demonstrate the principle.

While the encoder design is modular, each ANF model must be associated with a defined place along the tonotopic axis — i.e., a mapping to a characteristic frequency. This is essential for constructing the neurogram as a spatio-temporal representation of neural activity. The neurogram, denoted by $\widetilde{N}$, is a two-dimensional matrix, where each element $\widetilde{N}[t_k, f_i]$ captures activity for a specific time-frequency bin, derived from the spike trains of multiple fibers averaged across multiple repetitions. Here, we generate the neurogram from raw spike trains in two steps, binning and filtering.

3.1.1. Binning

After being presented with a stimulus, each modelled ANF produces a spike train, a sequence of discrete action potentials over time. This is repeated over k repetitions for each of the m ANF models in the population. This produces a total of mk spike trains for each simulation, denoted by $s_i(t)$, where $i \in [0, \ldots, mk)$. To generate a time–frequency representation, the spike trains are discretized along both the temporal and frequency dimensions.

Temporally, the spike train is divided into fixed-width time bins of size Δt, yielding spike counts:

$$b_i[t_k] = \int_{t_k \Delta t}^{(t_k+1)\Delta t} s_i(t) dt, \tag{3}$$

where $b_i[t_k] \in \mathbb{N}$ represents the number of spikes for trial i in time bin t_k. An additional frequency binning step is performed for trials that share the same frequency bin. For every frequency band f_i, the spike counts from each associated trial are pooled:

$$\widetilde{N}_{f_i}[t_k] = \sum_{i \in \mathcal{T}_{f_i}} b_i[t_k], \tag{4}$$

where $\mathcal{T}_{f_i}$ is the set of all trials assigned to frequency band f_i. This results in a neurogram, a 2D matrix $\widetilde{N}[t_k, f_i]$, where each element captures the magnitude of neural activity for a given time and frequency band.

3.1.2. Filtering

The neurogram $\widetilde{N}$ is smoothed along the time axis using a symmetric Hann window to reduce temporal variability. For each frequency band f_i, the smoothed signal is computed as:

$$\widetilde{N}_{f_i}[t_k] = \frac{1}{\sum_m h[m]} \sum_m \widetilde{N}_{f_i}[t_k - m] \cdot h[m], \tag{5}$$

where $h[m]$ is a Hann window of length m, given by $h[m] = 0.5 - 0.5\cos(\frac{2\pi m}{M-1})$. This operation smooths the neurogram along the temporal axis while preserving its frequency resolution.

Scaling. The filtered neurogram is normalized by scaling all values $\widetilde{N}[t_k, f_i]$ to the range $[0, 1]$ based on the minimum and maximum across the entire matrix, yielding relative activity patterns.

3.2. Encoding Neurograms

The encoder was configured with 64 frequency bands, spaced on a Mel scale between 150 Hz and 10 500 Hz. The same configuration was used wherever possible for both modelling paradigms, i.e., NH and EH. While this is not strictly necessary, it simplifies the configuration and shows generalizability. For each frequency band, ten fibers were simulated, each using 20 independent trials, generating a total of 12,800 spike trains per stimulus condition. Stimuli were presented at 50 dB Root Mean Square (RMS) Sound Pressure Level (SPL)[3], and the binsize of the generated neurograms $\widetilde{N}$ was set to $\Delta t = 36\mu$s[4]. A Hann window of length $H = 1500$, which is $H \cdot \Delta t = 0.054$ s[5], was used for filtering. As mentioned in Section 3.1, to generate $\widetilde{N}$, each ANF model needs to be associated with a frequency bin f_i. This is explained in more detail in Section 3.3 for the NH model and in Section 3.4 for EH.

[3]This lower presentation level was chosen to avoid the non-linear behaviour the Bruce et al. (2018) model shows for louder stimuli. Especially under stimuli with noise conditions, this produces an always-on behaviour for the model (see Figure 8), which severely impacts the reconstruction quality.

[4]The same length as the stimulus waveform used for EH.

[5]A multiple of the cycle speed of the SCS, 1500 / 15 = 100 cycles.

3.3. Normal Hearing

We simulate spike trains under normal hearing (NH) conditions using the auditory nerve fiber (ANF) model developed by Bruce et al. (2018). The model includes an inner hair cell and synapse component that captures realistic peripheral encoding dynamics, including neural adaptation and refractoriness. For each frequency band, two low, two medium, and six high spontaneous rate fibers were simulated. Each fiber's characteristic frequency (CF) was set to the center frequency of its corresponding frequency band. For other parameters, default values have been used as provided by Zilany and Bruce (2023) with the synapse modifications from Bruce et al. (2023).

Using the procedure outlined in the previous sections, we generated an example neurogram for a stimulus containing bird song, shown in Figure 3A. From the figure, the relationship between the spectrogram (Figure 3B) and the neurogram (Figure 3C) is clearly visible. Note also that some information is lost, and that the neurogram has sharper transitions than the spectrogram. Additionally, even though the original signal has no noticeable frequency content below 2 000 Hz, the spontaneous spiking activity does cause the neurogram to have a signal for those frequencies.

3.4. Electrical Hearing

For EH, we use the modeling pipeline presented in Section 2.4. The used stimulus waveform is a biphasic cathodic-first square pulse with a phase width of 18 μs. The modelled geometry (HC3, see: Kalkman et al. (2014)) has a model equivalent of a HiFocus Mid-Scala cochlear implant, which has 16 electrode contacts. The fibers are modeled without neural degeneration (Kalkman et al., 2022), and a threshold profile for 3 200 fibers was generated (see Figure 4), spaced evenly throughout the cochlea. To accommodate the current steering, 135[6] virtual electrode channels were modeled by calculating the activation threshold for a fiber when stimulated simultaneously by two adjacent electrodes.

SpecRes operates at a sampling rate of 17 400 Hz, which means the Nyquist frequency of 8 700 Hz effectively cuts off higher frequency content

[6] 120 + 15 'duplicate' channels

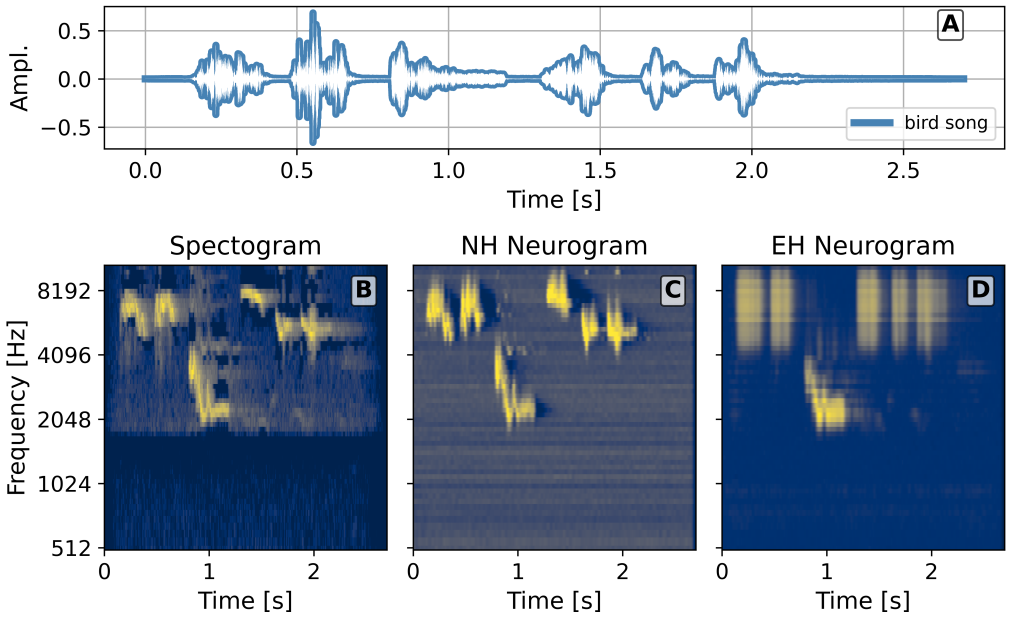

Figure 3: Examples of neurograms generated with the two different models. The top figure (A) shows the processed audio stimulus, which is a short fragment of a bird singing. The bottom left figure (B) shows a spectrogram of the stimulus, displayed using a mel scale. C shows the neurogram generated using the Bruce et al. (2018) model, Figure D shows a neurogram generated using the EH model described in Section 2.4. The color scale of the spectrogram ranges from 0 to -80 dB, and from 0 to 1 for the neurograms. Lighter colors indicate higher values.

from the audio signal. Additionally, the limited insertion depth of the implant means that low-frequency signals are also not correctly transferred to the CI user. Moreover, each electrode carries the band-pass filtered signal of a specific frequency band, which does not necessarily correspond to the tonotopic location of the electrode. This is illustrated in Figure 5, which shows the mismatch between the signal transmitted by each electrode contact and the tonotopic organization of the cochlea, as predicted by the Greenwood function. From the figure, it can be seen that the signal transmitted by the implant is generally around one octave lower than the tonotopic frequency of the neurons stimulated by that contact (Carlyon et al., 2010). This is one of the reasons that a CI can sound too high-pitched, especially for new users (Mertens et al., 2022). However, over time, neural adaptation can allow the brain to adapt to a new tonotopic map and reassign meaning to the

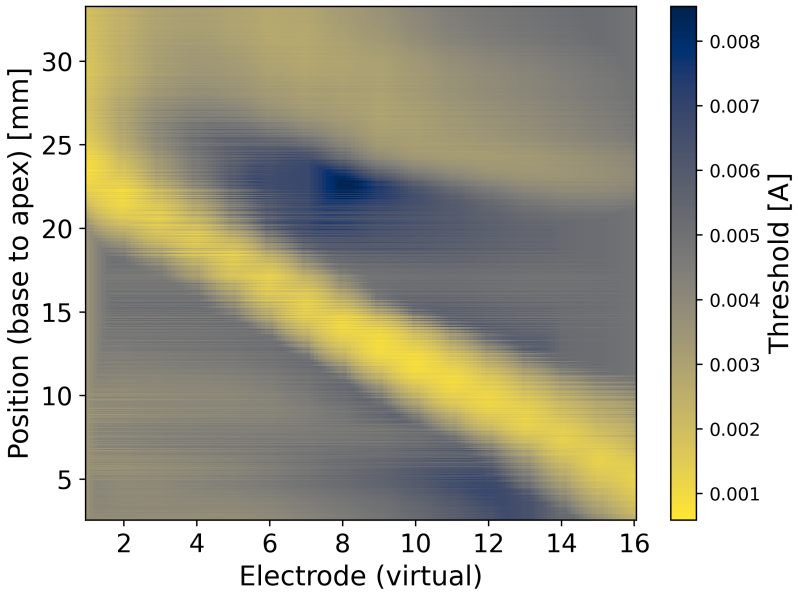

Figure 4: Heatmap visualization of a threshold profile specifically generated for SpecRes, containing 135 virtual electrode channels for 3 200 simulated auditory nerve fibers. The virtual channels are created between two stimulating electrodes, with nine evenly spaced steps. The color indicates the activation threshold of the fiber when stimulated by a given electrode pair.

frequencies, thereby normalizing pitch perception (Reiss et al., 2007). While this is not the case for all CI users, we take this as a given in assigning a frequency to a fiber. Specifically, we remap the original Greenwood frequencies to the electrode-specific operating frequencies used by SpecRes. This is what is shown by the orange line in Figure 5. We use interpolation to create a continuous frequency profile for all modeled fibers, smoothly transitioning from the natural Greenwood frequencies to the 'learned' frequency-place assignments dictated by the implant's electrode configuration.

Based on this 'learned' frequency mapping, we randomly select ten fibers for each frequency band f_i that have a frequency mapping falling within that band. We simulate with the parameters of PHAST+ as specified by the 'Average Fiber' in de Nobel et al. (2024).

Figure 3D shows the neurogram generated by the EH model in response to

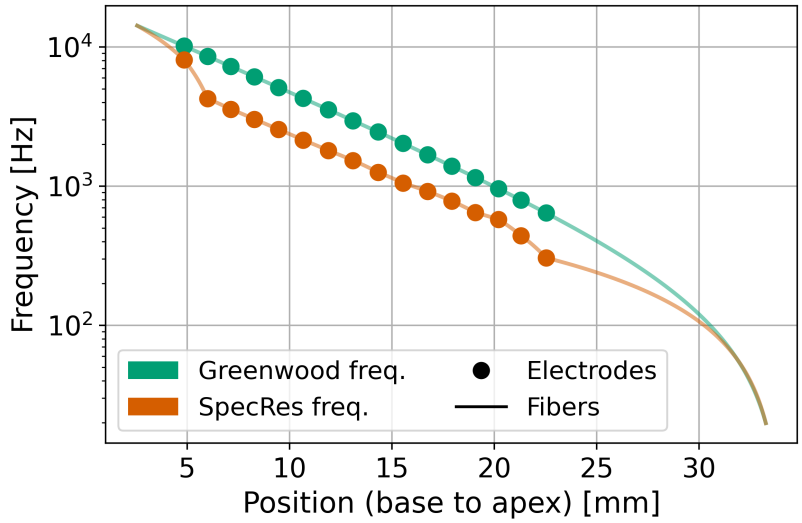

Figure 5: Line plot showing the frequency to cochlear position relation for both the stimulating electrodes and the individual fibers. The fibers are shown as a solid line, while the electrodes are visualized using dots. The positional frequency mapping, based on the Greenwood function, is shown in green. The frequency mapping, as used by SpecRes, is shown in orange, with the fiber frequency linearly interpolated to the operating frequencies of the electrodes.

the same stimulus containing bird song. From the figure, it is clear that while the frequency alignment of the model is appropriate, due to the 'learned' frequencies, it has a much lower temporal precision than the NH neurogram. This is partly due to CIS, which requires that all electrode pairs be stimulated for a single cycle of the strategy. Moreover, since there is only one electrode pair that stimulates signals over 4 248 Hz (see Table 2 in the Appendix), there is very little precision for high-frequency stimuli.

3.5. Decoding Neural Responses

The encoder evaluates an arbitrary simulation model and generates a binned and filtered neurogram $\widetilde{N}$. The decoder component of NeuroVoc (see Figure 2) then performs three sequential operations to reconstruct a time-domain signal $\hat{x}[t]$. These operations include preprocessing, reconstruction, and postprocessing, and will be explained in more detail below.

3.5.1. Preprocessing

To reduce the computational cost of the reconstruction stage, the neurogram is first downsampled along the temporal axis using polyphase filtering. If we have the original neurogram $\widetilde{N}$, which consists of a total of N time frames, we first compute the target number of time frames $n_s = \lceil \frac{N}{32} \rceil$. Here, 32 represents a fixed hop size, which is the number of frames to skip[7]. Resampling was performed with a rational factor $\frac{n_s}{N}$, reduced to its lowest terms, and included an anti-aliasing low-pass filter to minimize spectral distortion.

After resampling, each neurogram value $\widetilde{N}[t_k, f_i]$ was clipped to ensure all values remained within $[0, 1]$, and rescaled to a decibel-like range using a linear mapping:

$$\widetilde{N}[t_k, f_i] = -80 + 80 \times \min\left(1, \max\left(0, \widetilde{N}[t_k, f_i]\right)\right), \tag{6}$$

preserving the relative magnitudes. A floor of -80 dB relative to full scale (0 dB) is imposed to suppress irrelevant low-energy content. This threshold corresponds to the default dynamic range in standard signal processing toolkits (Brian McFee et al., 2015). Finally, the decibel-scaled values were converted to a power scale relative to a 50 dB reference, according to:

$$\widetilde{\widetilde{N}}[t_k, f_i] = 50.0 \times 10^{\frac{1}{10}\widetilde{N}[t_k, f_i]} \tag{7}$$

This transformation yields a representation analogous to a Mel-band spectrogram, i.e., power in Mel bands over time, serving as the input to the reconstruction stage.

3.5.2. Reconstruction

The goal of the reconstruction stage is to recover a time-domain waveform from the processed neurogram representation. Given a rescaled and downsampled neurogram $\widetilde{N}$, the first step in this process is to retrieve a magnitude spectrogram with a linear frequency scale (see Section 2.1). Currently, the frequency bins of $\widetilde{N}$ are on a Mel scale, and for the signal reconstruction stage, we require it to use the same scaling as an STFT.

[7] The same hop size that is used in the reconstruction stage

To accomplish this, we construct a Mel filterbank $\mathcal{M}$, which defines a linear transformation, i.e., a 2D matrix, from FFT bins to Mel-frequency bins. The filterbank maps 512-point FFTs to the frequency scale used in the encoder. To estimate the underlying STFT power spectrum, we solve a non-negative least squares (NNLS) problem:

$$\arg\min_{\hat{S}^2 \geq 0} \|\mathcal{M}\hat{S}^2 - \widetilde{N}\|_F,$$

where $\hat{S}^2$ denotes the estimated power spectrum and $\|\cdot\|_F$ the Frobenius norm. The resulting estimate is then converted to a magnitude scaling by taking the elementwise square root.

Griffin Lim. After estimating the STFT amplitude spectrum $\hat{S}$, the final step is to reconstruct a time-domain waveform $\hat{x}[t]$. Since phase information is not available in the generated magnitude spectrum, phase reconstruction is performed using the Griffin-Lim algorithm, as described in Section 2.2. Here, the version proposed by Perraudin et al. (2013) was used, and the algorithm was configured with a 512-point Hann window, executed for 320 iterations. A small hop size of 32 samples was chosen relative to the 512-sample window. This increases the redundancy between each consecutive frame and enhances the stability of the algorithm. The hop size matches the downsampling factor applied earlier to the neurogram, ensuring that the reconstructed waveform $\hat{x}[t]$ has a sampling rate consistent with the neurogram, namely $1/\Delta t$.

3.5.3. Postprocessing

In the final stage of the pipeline, the reconstructed waveforms $\hat{x}[t]$ are resampled to the original sampling frequency of the input signal f_s. Since the signal is periodic, Fourier-based resampling is used. Finally, $\hat{x}[t]$ is scaled to 50 dB RMS SPL, producing a reconstructed signal with the same amplitude scaling as the input (see Section 3.2).

4. Experiments

We perform two experiments to validate whether the proposed approach provides satisfactory reconstructions. First, we examine the spectrograms for a short speech segment and compare the unprocessed sound with the

reconstructed sounds using both models (Section 4.1). Secondly, to asses the intelligibility and perceptual quality of the vocoder, we perform an online Digits-in-Noise test, to investigate the synthesized audio from a (normal-hearing) listeners perspective (Sections 4.2 & 4.3).

4.1. 'Choice'

We qualitatively examine the reconstructed signals for the word *choice* in the first experiment. For both the normal hearing (NH) and electrical hearing (EH) models, sounds were reconstructed using the approach outlined in the previous section.

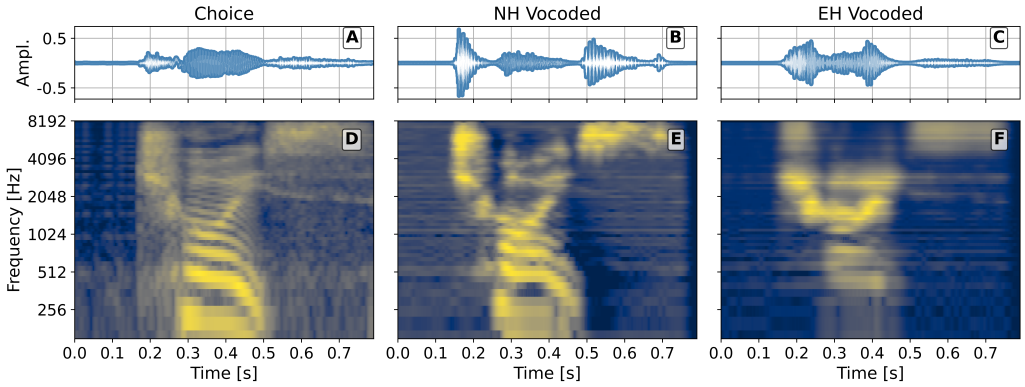

Figure 6: Spectrogram and waveform visualizations for the word 'choice'. The leftmost (A) panel shows the unprocessed sound. Reconstructed sounds are shown for both the normal hearing (NH) vocoder (B) and the electrical hearing (EH) vocoder (C). Spectrograms, shown in Figures D-E-F, are generated by applying an STFT to the (reconstructed) signal for a 512-point FFT, displaying the magnitude spectrum in a dB scale, ranging from -80 to 0, with lighter colors indicating higher energy.

In the top panel of Figure 6 (A-C), the reconstructed waveforms are shown alongside the original input stimulus. We observe that while the timing of the reconstructed signals is well aligned with the original, the amplitude is not. Interestingly, the amplitude of the EH reconstruction (Figure 6B) more closely resembles that of the original signal compared to the NH reconstruction (Figure 6C). From what we have observed, this is partly due to the response characteristics of the auditory nerve fiber (ANF) model by Bruce et al. (2018), which is sensitive to stimulus onsets following silence.

Specifically, even low-intensity inputs can trigger spikes after a period without stimulation. Since perceived loudness in our framework is based on the number of simultaneous spikes within a time-frequency bin, this results in a prominent peak in the reconstructed signal at the onset for NH, even when the input amplitude is relatively low. A similar effect is present in the EH model, but it is less pronounced.

When we shift our focus to the (reconstructed) signal's frequency content over time—illustrated in the spectrograms in the bottom row of Figure 6—a different picture emerges. Here, the NH vocoder performs remarkably well: it preserves most of the harmonic and spectral content of the input signal, despite the amplitude mismatch observed in the waveform domain. The fundamental frequency and its harmonics are clearly visible and correctly aligned in time and frequency. In contrast, the EH vocoder exhibits substantial spectral degradation. Much of this degradation can be attributed to the limited bandwidth and the coarser frequency binning of the SCS in the CI model. This is especially evident for higher frequencies, transmitted by a single electrode contact, which causes a smearing in the spectrogram. This could also be observed in Figure 3D. In addition, channel interaction, caused by current spread in the cochlea, further degrades frequency selectivity. Because electrical stimulation from one electrode can spread and activate adjacent auditory nerve regions, the effective independence between channels is reduced, leading to overlapping neural excitation patterns and a blurring of spectral details.

These differences in reconstruction quality are not unexpected. The NH model is designed to represent the peripheral encoding of sound in a healthy auditory system, while the EH model approximates a CI user. As such, the degraded spectral fidelity observed in the EH reconstructions aligns with each model's intended use: CI users often perceive sound with reduced clarity and resolution compared to normal hearing individuals (Bonham and Litvak, 2008). Therefore, the vocoder results shown here are consistent with the perceptual limitations imposed by the underlying models.

Adding noise to 'choice'. When speech-shaped noise is added to the *choice* stimulus at an SNR of –4 dB, the resulting reconstructions are shown in Figure 7(A-C). The mismatch in amplitude between the original and reconstructed signals persists for both the NH and EH models. Moreover, we see

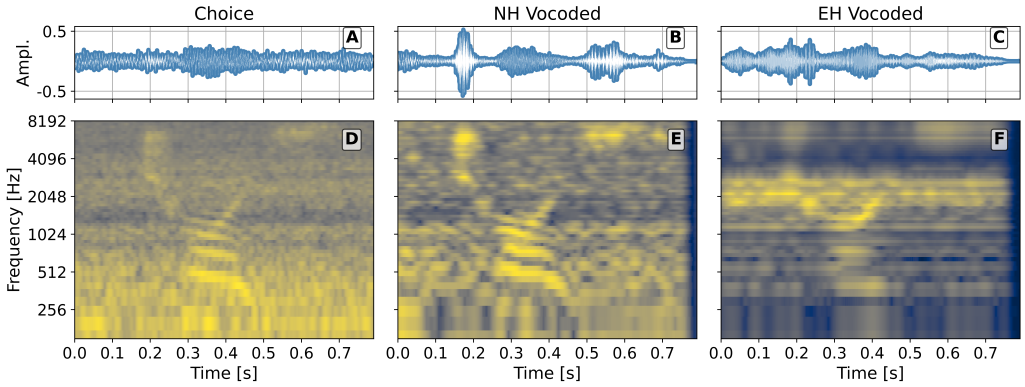

Figure 7: Spectrogram and waveform visualizations for the word 'choice', mixed with speech-shaped noise at -4 dB SNR. The leftmost panel (A) shows the unprocessed sound. Reconstructed sounds are shown for both the normal hearing (NH) vocoder (B) and the electrical hearing (EH) vocoder (C). Spectrograms, shown in Figures D-E-F, are generated by applying an STFT to the (reconstructed) signal for a 512-point FFT, displaying the magnitude spectrum in a dB scale, ranging from -80 to 0, with lighter colors indicating higher energy.

a clear difference if we compare the amplitude at the beginning of the signal with the amplitude at the end of the reconstructed signal, which both should only contain noise. Specifically, the EH vocoder produces a much larger amplitude at the onset than the (input) signal strength alone would suggest.

Turning to the spectrograms, several interesting observations can be made. For the NH vocoder, the structure of the original speech signal remains relatively well preserved despite the added noise. However, how the auditory nerve model by Bruce et al. (2018) encodes the noise introduces distinct distortions. While the input noise exhibits a relatively flat spectral profile — i.e., consistent energy across frequencies — the reconstructed spectrogram shows irregular "clumping" in intensity. Specifically, the energy fluctuates in bursts, alternating between high and low amplitudes over time. This pattern can be attributed to the refractory properties of the auditory nerve fibers. Because each fiber has a recovery period following an action potential, it cannot respond uniformly to a constant or broadband input such as noise. As a result, sustained stimuli like noise are encoded in a temporally modulated way. This behavior is clearly visible in Figure 8, where a zoomed version of the neurogram shows periodic activity interspersed with silent intervals.

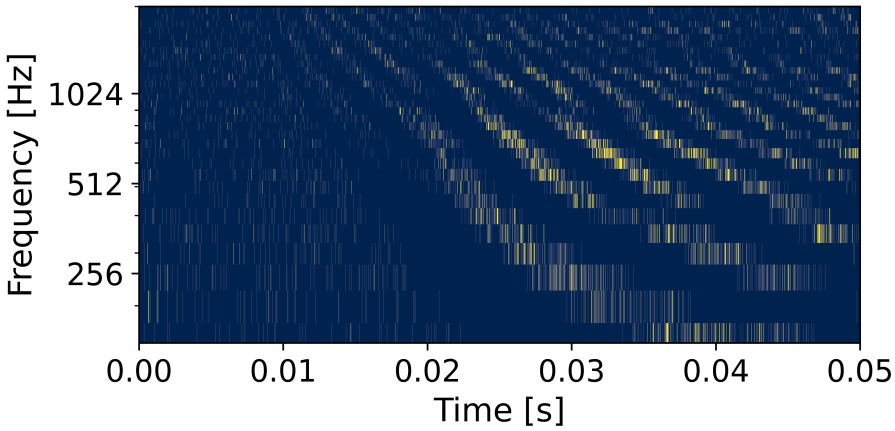

Figure 8: Neurogram of the Bruce et al. (2018) model for the stimulus 'choice', mixed with speech-shaped noise at -4 dB SNR. The first 0.05 seconds of the neurogram are shown, for the fibers with a CF $\in [150, 2000]$ Hz.

In contrast, this refractory-driven modulation is less evident in the EH vocoder (Figure 7F). However, the impact of noise manifests differently: channel interaction becomes substantially more pronounced. Specifically, the spectral smearing in the mid-frequency range (approximately 1 000–3 000 Hz) increases, causing certain frequency bands to become overemphasized. This leads to a suppression of finer spectral details and a loss of clarity in the reconstructed signal.

Overall, the added noise has a more detrimental effect on the EH model than on the NH model. This is consistent with real-world observations: cochlear implant (CI) users are generally more affected by noisy environments than normal-hearing listeners (Cullington and Zeng, 2008). The vocoder reconstructions mirror this limitation, reinforcing that the EH model captures key perceptual challenges CI users face.

4.2. Digits in Noise

This section evaluates two neural vocoders using the Digits-in-Noise (DIN) test (Smits et al., 2013). The test is based on Dutch speech material consisting of 120 digit triplets. It was conducted online with normal-hearing listeners, each of whom completed three test conditions: the standard DIN

test (unprocessed), the test using the NH vocoder, and the test using the EH vocoder. Further details on the test procedure can be found in Section 4.3. To begin, we provide an overview of the test data by comparing key statistics of the reconstructed signals to those of the original digit triplets in the next section.

4.2.1. Statistics

In this section, we quantitatively evaluate the neural vocoders on the 120 clean (noiseless) speech stimuli from the Dutch DIN test. To ensure consistent measurements, all audio files—both the original and reconstructed signals—are amplitude-normalized to –20 dB relative to full scale (FS). Before comparison, the reconstructed signals are temporally aligned with their corresponding input stimuli. This is necessary because the neural vocoder introduces a slight delay: ANFs respond only after a stimulus occurs, causing a small timing offset. We address this by applying Dynamic Time Warping (DTW) (Berndt and Clifford, 1994), which non-linearly aligns each reconstructed signal with its original. After alignment, we evaluate the reconstructions using two objective measures:

1. **Mean Square Error (MSE)** between the input waveform $x[t]$ and the reconstructed waveform $\hat{x}[t]$, defined as:

$$\mathrm{MSE} = \frac{1}{T}\sum_{t=1}^{T}\left(x[t]-\hat{x}[t]\right)^2 \tag{8}$$

   This metric quantifies amplitude deviations and reflects how well the reconstructed waveform preserves the dynamic range of the original signal.

2. **Mel-Cepstral Distortion (MCD)**, which compares the mel-cepstral coefficient (MCC) sequences of the original and reconstructed signals. MCC sequences represent the spectral envelope of a sound signal. They are calculated by applying a discrete cosine transform to the log-scaled power spectrum produced by a Fourier transform mapped onto a mel frequency scale. Given c_t and $\hat{c}_t$ as the MCC sequences at frame t, for the original and reconstructed signal, respectively, MCD is defined as:

$$\mathrm{MCD} = \frac{10}{\ln 10}\sqrt{2\sum_{m=1}^{M}\left(c_t^{(m)}-\hat{c}_t^{(m)}\right)^2} \tag{9}$$

where $M = 13$ is the number of mel-cepstral coefficients. MCD is commonly used to assess the quality of parametric speech synthesis systems (Kominek et al., 2008). A lower MCD indicates that the synthesized mel-cepstral sequence closely matches that of the original speech, suggesting higher perceptual similarity between the synthetic and original signals.

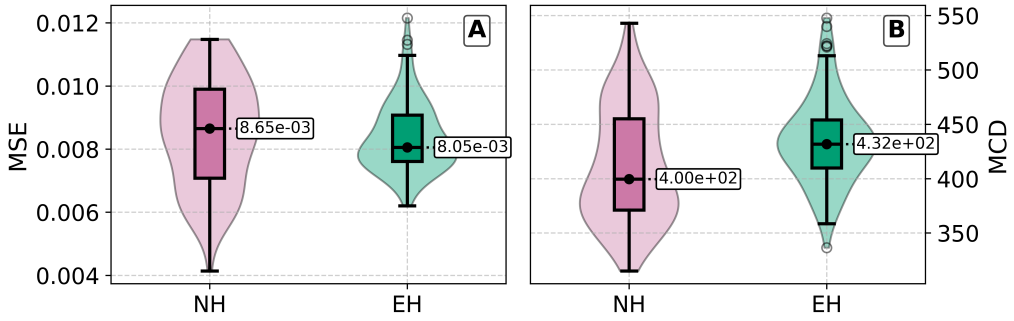

Figure 9: Summary statistics of the reconstructed audio samples of the (noiseless) speech material of the digits-in-noise test compared against the original samples. The left panel (A) shows the Mean Square Error (MSE) for both the Normal Hearing (NH) and the Electrical Hearing (EH) vocoder. The right panel (B) shows the Mel-Cepstral Distortion (MCD) for both models. For both, lower is better.

Together, these metrics provide complementary insights into vocoder performance, similar to the visual analysis presented in the previous section. Specifically, they capture both the temporal amplitude structure via MSE and the preservation of spectral content via MCD. Figure 9 presents a boxplot summarizing the results.

The patterns observed are consistent with those found in the analysis of the 'choice' stimulus (see Section 4.1). For the NH vocoder, the reconstructed waveforms exhibit greater variability in relative amplitude, as reflected by higher MSE values and a larger standard deviation, shown in Figure 9A. In contrast, the EH vocoder shows more stable amplitude reconstruction.

However, the opposite trend is observed in the spectral domain, as displayed in Figure 9B. The NH vocoder yields lower mel-cepstral distortion (MCD), indicating superior preservation of the original stimuli's spectral fea-

tures. By comparison, the EH vocoder shows greater spectral degradation.

These differences between the NH and EH vocoders are statistically significant. A two-sided Mann–Whitney U test yielded p-values of 0.0197 for MSE and 0.0009 for MCD, confirming that the vocoders differ meaningfully on both temporal and spectral reconstruction metrics for noiseless speech samples, given a confidence bound $\alpha = 0.05$.

4.3. Online Digits in Noise Test

In this section, we describe the experimental setup used to evaluate the intelligibility of vocoder-reconstructed speech using the Digits-in-Noise (DIN) test. A custom web-based testing platform was developed to administer the test following a standardized, adaptive two-up two-down procedure (Smits et al., 2013). Each participant completed three versions of the test: one using unprocessed stimuli (standard DIN), one using speech reconstructed by the NH vocoder, and one using speech reconstructed by the EH vocoder. The following subsections provide a detailed description of the stimulus preparation, test procedure, and study population.

4.3.1. Stimulus Preparation

Each of the 120 Dutch digit triplets used in the Digits-in-Noise (DIN) test was mixed with speech-shaped noise across a range of signal-to-noise ratios (SNRs) to generate the test materials. Every triplet was mixed with a randomly sampled noise instance at SNRs ranging from –20 dB to +10 dB in 2 dB increments, resulting in 16 SNR conditions per digit triplet. This yielded a total of 1,920 noisy speech signals. These signals formed the unprocessed (raw) stimulus set. The same set was then processed through the NH and EH vocoder pipelines, resulting in two additional vocoded stimulus sets —one for each model —yielding three distinct corpora of noisy speech. Mixing and vocoding were performed offline before test deployment, resulting in a consistent body of test stimuli for all users. The resulting stimuli were amplitude-normalized to –20 dB full scale (FS) to control the presentation loudness.

4.3.2. Procedure

The DIN test was implemented on a custom-built website, allowing participants to complete the task remotely using their own devices and headphones. The procedure started with a calibration step adapted from Shehabi

et al. (2025), in which participants adjusted their device volume based on two sentences played 25 dB apart in RMS level: one intended to be clearly audible, and the other loud but not uncomfortable. All subsequent stimuli were presented diotically at a fixed level (–20 dB FS), 5 dB below the high-level sentence and 20 dB above the low-level sentence, ensuring audibility even at the lowest SNR of –20 dB.

Following calibration, participants completed a single practice trial to familiarize themselves with the task. The interface was simple and consistent across all test conditions. Each participant completed three DIN tests in randomized order: unprocessed speech (standard DIN), NH-vocoded speech, and EH-vocoded speech. Each test consisted of 24 digit-triplet presentations. For each trial, a stimulus was randomly sampled from the 120-triplet corpus at the current SNR. Each trial began when the participant clicked a single-use playback button, which played the audio. After listening, they selected their response using on-screen digit buttons, with the option to revise their answer before submission.

An adaptive two-up two-down procedure was used to vary the SNR based on response accuracy. All tests started at an initial SNR of 0 dB, with values bounded between –20 dB and +10 dB, and were increased by two dB on a correct answer (all digits correct) and decreased otherwise. Performance was quantified using the speech reception threshold (SRT), following the protocol by Smits et al. (2013), defined as the average SNR of presentations 5 through 25. The SNR of the 25th presentation is the hypothetical level of the presentation after the last presentation, based on the final adaptive step.

All participant data was collected anonymously. Only age, whether participants believed they had normal hearing, and whether they had previously completed the DIN test were recorded.

4.3.3. Study population

A total of 55 participants with self-reported normal hearing completed the study. All participants were fluent in Dutch and completed the test remotely using their own devices and headphones. Data was collected anonymously through the web platform, and no personally identifiable information was recorded. Three participants were excluded from the dataset. Two participants did not complete all three lists, and the other scored an SRT of -5.5

dB on the unprocessed test, which was deemed too high an outlier for NH.

4.3.4. Results

The goal of this experiment was to evaluate how well the neural vocoders preserve speech intelligibility (SI) in noise, as measured by the Digits-in-Noise (DIN) test. We hypothesized that: (1) the vocoded conditions would perform in line with expected differences between normal hearing and cochlear implant (CI) listeners; (2) added noise would have a more detrimental effect on the EH (electrical hearing) vocoder compared to the NH (normal hearing) vocoder; and (3) although both vocoders would introduce some degradation, performance in the NH condition would more closely resemble that of the unprocessed (raw) speech condition.

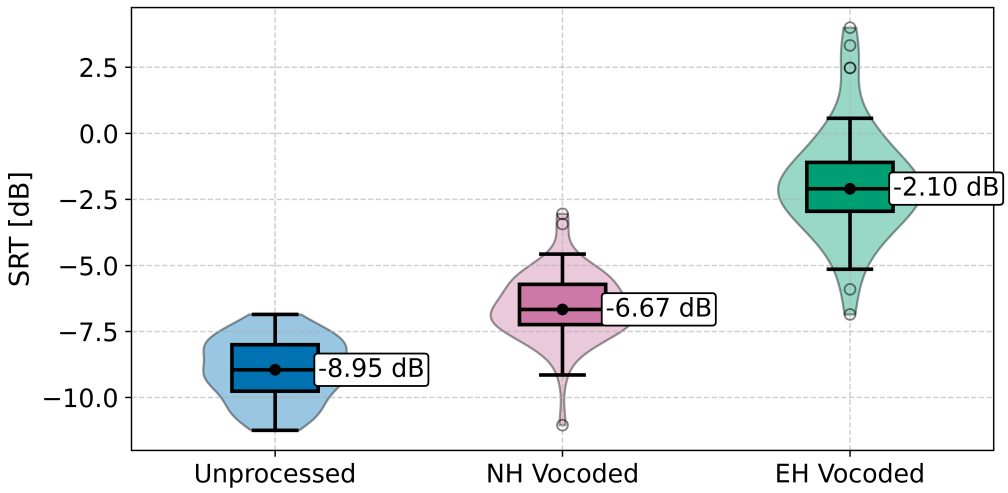

Figure 10: Box plot of the speech reception threshold (SRT) for the DIN test. Three test conditions are shown: unprocessed (normal DIN), NH vocoded sound, and EH vocoded sound. The median SRT is provided as an annotation to the plot.

Figure 10 shows a boxplot of the speech reception thresholds (SRTs) across the three test conditions: unprocessed, NH vocoded, and EH vocoded. *The results are consistent with our hypotheses.* All groups are statistically different from each other (tested by a Welch's t-test, $\max p \approx 2.5 \cdot 10^{-16}$, $\alpha = 0.05$). The unprocessed condition, i.e., the standard test, yielded the lowest SRTs, indicating the highest intelligibility. The EH vocoded condition significantly elevated the SRT of the participants by 7.12 dB on average.

The NH vocoded condition produced intermediate results, with an elevated SRT of 2.4 dB, suggesting better preservation of speech cues in noise but still measurable loss compared to the unprocessed condition.

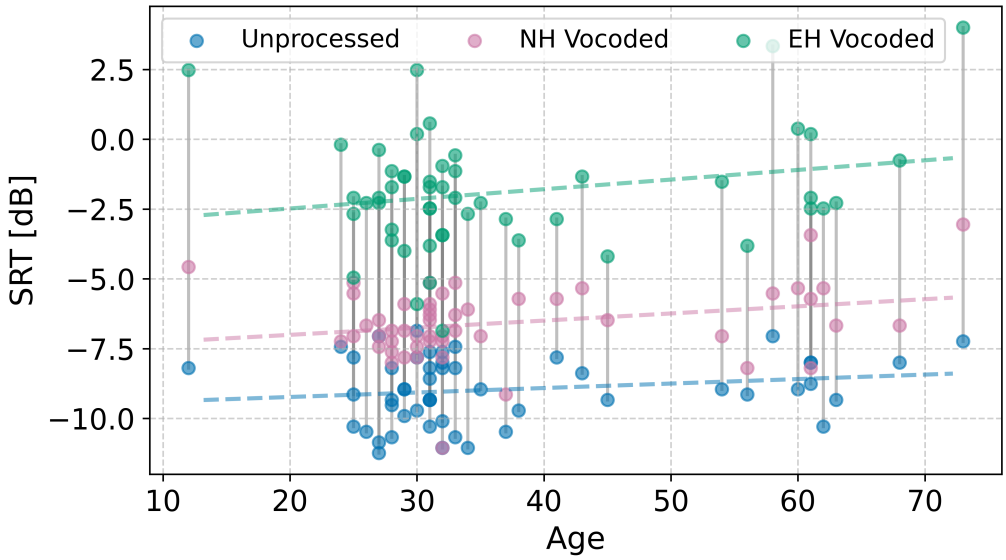

Figure 11: Speech reception threshold (SRT) for the DIN test for each of the three test conditions, unprocessed (normal DIN), NH vocoded sound, EH vocoded sound, shown as a function of participant age. A faded grey line connects the test results for each participant across the three conditions.

Figure 11 shows the effect of age on the test results. There is a minor correlation ($R^2 \approx 0.011$) between age and the DIN test SRT, which aligns with the literature, indicating a negative association between age and SI in noise (Goossens et al., 2017). Between test conditions, there is only a minor difference in the slope of age vs. SRT, with the slope for the EH vocoded group being the steepest.

5. Discussion

In this study, we introduced *NeuroVoc*, a model-agnostic vocoder framework capable of reconstructing acoustic waveforms from simulated neural activity. Unlike recent approaches that rely on machine learning or data-driven methods (Park et al., 2023; Daly, 2023), our system uses classical

signal processing techniques to generate intelligible speech. Its simplicity and modularity allow it to interface with arbitrary ANF models without requiring specialized adaptation. The only requirement of the method is a consistent way of generating a neurogram, which requires a mapping from fiber to frequency. This enables the evaluation of different experimental conditions across diverse modeling paradigms while keeping the vocoder consistent. Such flexibility is particularly valuable for CI research, where specific speech coding strategies often require custom vocoding implementations (Cychosz et al., 2024). With our method, these strategies can be evaluated within a unified framework. For example, one could directly compare the SpecRes strategy used in this study to the commonly used ACE strategy, from a different CI manufacturer. Moreover, as demonstrated here, the framework can accommodate entirely different auditory models, such as the normal-hearing model of Bruce et al. (2018) and the electrical hearing model present in Section 2.4, within the same computational pipeline. Importantly, our method requires minimal parameter tuning. We used default settings for the auditory models and selected vocoder parameters based on generalizability rather than dataset-specific optimization. Despite this, the system performs robustly, indicating that the reconstruction method from neurograms is effective even without fine-tuning.

5.1. Reconstruction Quality

Our results show that the reconstruction quality aligns with the characteristics of the underlying auditory models from which the neurograms were generated. In the normal hearing condition, the reconstructed waveforms were generally of higher quality. Although the amplitude dynamics were somewhat unstable—likely due to pronounced onset responses following periods of silence—the NH vocoder preserved spectral structure well. Harmonic content was clearly represented, resulting in reconstructions with rich frequency detail. Notably, in response to constant stimuli such as noise, the NH model does not produce a steady output due to the refractory behavior of the simulated auditory nerve fibers. This results in temporally "clumped" neural activity, which translates into amplitude fluctuations in the reconstructed waveform and reduces the perceived continuity and quality of the sound.

The cochlear implant (CI) condition, modeled by the electrical hearing (EH) paradigm, exhibited much reduced temporal and spectral specificity.

The limited frequency resolution imposed by the implant and the speech coding strategy used was reflected in the degraded reconstructions. Additionally, spectral smearing, especially under noisy conditions, had a detrimental effect, diminishing signal clarity in affected frequency regions. These results are consistent with known perceptual limitations experienced by CI users, suggesting that the vocoder accurately reproduces the characteristic degradations associated with electrical hearing (Shannon et al., 1995; Mertens et al., 2022).

It should also be noted that while many studies (Johannesen et al., 2022; Leclère et al., 2023; Gajecki and Nogueira, 2022) use amplitude-based distance measures such as MSE to evaluate reconstruction quality, this does not necessarily measure intelligibility. For example, as could be observed from Figure 9A, even though the MSE for the EH-vocoder was generally lower than that of the NH-vocoder, the severe spectral distortion imposed by the EH-model caused the reconstructed speech to be much less intelligible, as demonstrated by the results presented in Section 4.3.4.

5.2. Perceptual Evaluation

Behavioral testing using an online Digits-in-Noise (DIN) test further validated our framework. Participants were presented with unprocessed, NH-vocoded, and EH-vocoded speech stimuli. While each participant only performed a single trial for each of the three test conditions, the learning effect was mitigated over the entire population by randomizing the order of the tests. The relatively large study population strengthens the validity of the results. Participants' SRTs were elevated by approximately 7.1 dB on average for the EH vocoder relative to the unprocessed condition. The NH vocoder condition resulted in only a moderate SRT shift (2.4 dB), indicating that while the vocoder introduces some signal degradation, it still provided for a good reconstruction, even in noisy conditions.

5.2.1. Comparing against clinical data

We observe that our results are similar when compared to published clinical data for the Dutch digit in noise test, as displayed in Table 1. There is considerable variability between studies, especially for the groups with a CI. Our results for the unprocessed test are right within the middle of the

studies, with a mean value of -8.9 dB. If we pool all the data from the studies together and perform two one-sided Welch's t-tests (TOST), assuming unequal variance, and set the equivalence margin to half the reported SD, i.e. $\frac{0.7}{2} = 0.35$ dB, the results from our unprocessed data are statistically equivalent ($\max p \approx 0.04,\ \alpha = 0.05$).

Table 1: Overview of clinical Dutch DIN SRT scores reported in the literature. Data for de Graaff et al. (2016) was estimated from Fig. 1 (discont. noise, retest). The data for Vroegop et al. (2021) included only children (average age 11.8 ±3.6), the standard deviation was estimated from Fig. 5. Aggregated values for the mean and standard deviation per group calculated as: $\sum(n_i \cdot \bar{x}_i)/\sum n_i$ and $\sum((n_i-1)\cdot s_i^2)/\sum(n_i-1)$, where $\bar{x}_i$ and s_i^2 are the reported mean and std. dev. The results from our study have been included for the unprocessed and CI-vocoded groups (see Figure 10).

	NH			CI		
Study	n	Mean [dB]	SD	n	Mean [dB]	SD
Smits et al. (2013)	23	-8.8	0.6	-	-	-
Smits et al. (2016)	16	-9.3	0.7	-	-	-
de Graaff et al. (2016)	12	-9.5	1.0	16	-3.6	1.7
Kaandorp et al. (2015)	12	-9.3	0.7	24	-1.8	2.7
Stronks et al. (2025)	18	-8.4	0.6	18	-1.5	2.5
Vroegop et al. (2021)	-	-	-	58	-1.4	3.8
Aggregated	81	-9.0	0.7	116	-1.8	3.1
This study	52	-8.9	1.2	52	-1.9	2.1

Similarly, our results from the EH vocoded test are very close to the average SRT reported in the clinical studies, which are -1.9 dB and -1.8 dB, respectively. Applying the same TOST procedure, using an equivalence margin of $\frac{3.1}{2} = 1.51$ dB, indicates that the data for the EH-vocoded group is statistically equivalent ($\max p \approx 0.0002,\ \alpha = 0.05$) to the clinically reported DIN test scores of CI users. We should note the high subject-level variability within the CI group, which is represented in our test data, with the highest SRT variance found within the EH vocoded group.

These findings also reinforce the validity of administering the DIN test online. Shehabi et al. (2025); de Graaff et al. (2016) have shown that DIN results collected remotely under controlled conditions (e.g., with proper cali-

bration and headphone use) do not differ significantly from those obtained in clinical settings. The fact that our unprocessed test aligns with the clinically collected data summarized in Table 1 suggests that the observed performance differences across vocoder conditions are robust and not an artifact of the online testing environment.

Taken together, these results demonstrate that NeuroVoc provides an effective simulation tool for comparing auditory perception across various hearing conditions.

5.3. Future work

While the current study shows that our vocoder method provides realistic reconstructions and captures the characteristics of the used model, several avenues remain open for future work:

- Further model-based testing: The framework is well-suited for exploring alternative CI coding strategies. Exploring the impact of different strategies or strategy parameters within the vocoder framework presented here would be a logical next step, and could help easily prototype new methods.
- Comparing against standard vocoders: It would be interesting to see how our method compares against vocoders designed explicitly for a given implant/SCS.
- In the current study, the parameters of the method, e.g., number of Mel bands, smoothing filter size, and number of FFT components, are not fully explored. Future work might help uncover more suitable values that are potentially model-specific.
- Refractory behavior: The refractoriness in the ANF models limits the ability to encode constant or sustained stimuli. Better spatiotemporal smoothing techniques might help overcome unwarranted temporal modulation in the reconstructed sounds.
- Loudness modeling: Our system currently handles relative loudness only, normalized to a suitable range. Incorporating a more accurate loudness scaling based on the firing behavior of the modeled fibers could make the reconstruction more realistic.

6. Conclusion

Using classic signal processing techniques, we presented a flexible vocoder framework that reconstructs sound from simulated auditory nerve activity. The system supports arbitrary auditory models, enabling direct comparisons between normal hearing and cochlear implant conditions without requiring model-specific vocoders. Our results show that the vocoder captures characteristic differences between models and that reconstructed speech is intelligible, with perceptual performance aligning closely with clinical benchmarks. These findings demonstrate the framework's utility as a lightweight, interpretable tool for prototyping and evaluating auditory perception modelling experiments.

Acknowledgements

This collaboration project is co-funded by the PPP Allowance made available by Health Holland (grant LSHM20101), Top Sector Life Sciences & Health, to stimulate public-private partnerships. In addition, financial support was also provided by Advanced Bionics Corporation.

References

Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., and Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Scientific reports*, 9(1):874.

Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238.

Bäck, T. H., Kononova, A. V., van Stein, B., Wang, H., Antonov, K. A., Kalkreuth, R. T., de Nobel, J., Vermetten, D., de Winter, R., and Ye, F. (2023). Evolutionary algorithms for parameter optimization—thirty years later. *Evolutionary Computation*, 31(2):81–122.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:.

Blamey, P., Artieres, F., Başkent, D., Bergeron, F., Beynon, A., Burke, E., Dillier, N., Dowell, R., Fraysse, B., Gallégo, S., et al. (2012). Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: an update with 2251 patients. *Audiology and Neurotology*, 18(1):36–47.

Bonham, B. H. and Litvak, L. M. (2008). Current focusing and steering: modeling, physiology, and psychophysics. *Hearing research*, 242(1-2):141–153.

Briaire, J. J. and Frijns, J. H. M. (2000). 3d mesh generation to solve the electrical volume conduction problem in the implanted inner ear. *Simulation Practice and Theory*, 8(1):57–73.

Briaire, J. J. and Frijns, J. H. M. (2005). Unraveling the electrically evoked compound action potential. *Hearing Research*, 205(1):143–156.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto (2015). librosa: Audio and Music Signal Analysis in Python. In Kathryn Huff and James Bergstra, editors, *Proceedings of the 14th Python in Science Conference*, pages 18 – 24.

Bruce, I., Buller, A., and Zilany, M. S. A. (2023). Modeling of auditory nerve fiber input/output functions near threshold. In *Acoustics 2023*, Sydney, Australia. Conference poster.

Bruce, I., Irlicht, L., White, M., O'Leary, S., Dynes, S., Javel, E., and Clark, G. (1999). A stochastic model of the electrically stimulated auditory nerve: pulse-train response. *IEEE Transactions on Biomedical Engineering*, 46(6):630–637.

Bruce, I. C., Erfani, Y., and Zilany, M. S. (2018). 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. Computational models of the auditory system.

Carlyon, R. P., Macherey, O., Frijns, J. H., Axon, P. R., Kalkman, R. K., Boyle, P., Baguley, D. M., Briggs, J., Deeks, J. M., Briaire, J. J., et al. (2010). Pitch comparisons between electrical stimulation of a cochlear implant and acoustic stimuli presented to a normal-hearing contralateral ear. *Journal of the Association for Research in Otolaryngology*, 11:625–640.

Cullington, H. E. and Zeng, F.-G. (2008). Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. *The Journal of the Acoustical Society of America*, 123(1):450–461.

Cychosz, M., Winn, M. B., and Goupell, M. J. (2024). How to vocode: Using channel vocoders for cochlear-implant research. *The Journal of the Acoustical Society of America*, 155(4):2407–2437.

Daly, I. (2023). Neural decoding of music from the eeg. *Scientific Reports*, 13(1):624.

de Graaff, F., Huysmans, E., Qazi, O. u. R., Vanpoucke, F. J., Merkus, P., Goverts, S. T., and Smits, C. (2016). The development of remote speech recognition tests for adult cochlear implant users: The effect of presentation mode of the noise and a reliable method to deliver sound in home environments. *Audiology and Neurotology*, 21(Suppl. 1):48–54.

de Nobel, J., Martens, S. S., Briaire, J. J., Bäck, T. H., Kononova, A. V., and Frijns, J. H. (2024). Biophysics-inspired spike rate adaptation for compu-

tationally efficient phenomenological nerve modeling. *Hearing Research*, 447:109011.

Dekker, D. M. T., Briaire, J. J., and Frijns, J. H. M. (2014). The impact of internodal segmentation in biophysical nerve fiber models. *Journal of Computational Neuroscience*, 37(2):307–315.

Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, 11(2):169–177.

El Boghdady, N., Kegel, A., Lai, W. K., and Dillier, N. (2016). A neural-based vocoder implementation for evaluating cochlear implant coding strategies. *Hearing Research*, 333:136–149.

Frijns, J. H. M., Briaire, J. J., and Schoonhoven, R. (2000). Integrated use of volume conduction and neural models to simulate the response to cochlear implants. *Simulation Practice and Theory*, 8(1):75–97.

Frijns, J. H. M., de Snoo, S., and Schoonhoven, R. (1995). Potential distributions and neural excitation patterns in a rotationally symmetric model of the electrically stimulated cochlea. *Hearing Research*, 87(1):170–186.

Gajecki, T. and Nogueira, W. (2022). An end-to-end deep learning speech coding and denoising strategy for cochlear implants. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3109–3113.

Goossens, T., Vercammen, C., Wouters, J., and van Wieringen, A. (2017). Masked speech perception across the adult lifespan: Impact of age and hearing impairment. *Hearing Research*, 344:109–124.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

Hanekom, T. and Hanekom, J. J. (2016). Three-dimensional models of cochlear implants: A review of their development and how they could support management and maintenance of cochlear implant performance. *Network: Computation in Neural Systems*, 27(2-3):67–106.

Hines, A. and Harte, N. (2012). Speech intelligibility prediction using a neurogram similarity index measure. *Speech Communication*, 54(2):306–320.

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500.

jabeim (2025). Ab-generic-python-toolbox: A toolbox for simulating cochlear implant signal processing and the perception of sound by implant recipients. https://github.com/jabeim/AB-Generic-Python-Toolbox. Accessed: 2025-04-24.

Johannesen, P. T., Leclère, T., Wijetillake, A., Segovia-Martínez, M., and Lopez-Poveda, E. A. (2022). Modeling temporal information encoding by the population of fibers in the healthy and synaptopathic auditory nerve. *Hearing Research*, 426:108621.

Johnson, K. O. (2000). Neural coding. *Neuron*, 26(3):563–566.

Kaandorp, M. W., Smits, C., Merkus, P., Goverts, S. T., and Festen, J. M. (2015). Assessing speech recognition abilities with digits in noise in cochlear implant and hearing aid users. *International Journal of Audiology*, 54(1):48–57.

Kalkman, R. K., Briaire, J. J., Dekker, D. M., and Frijns, J. H. (2014). Place pitch versus electrode location in a realistic computational model of the implanted human cochlea. *Hearing research*, 315:10–24.

Kalkman, R. K., Briaire, J. J., Dekker, D. M., and Frijns, J. H. M. (2022). The relation between polarity sensitivity and neural degeneration in a computational model of cochlear implant stimulation. *Hearing Research*, 415:108413.

Kansaku, K. (2021). Neuroprosthetics in systems neuroscience and medicine. *Scientific Reports*, 11(1):5404.

Kiang, N. Y.-S., Watanabe, T., Thomas, E. C., and Clark, L. F. (1966). *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*, volume 35 of *MIT Research Monograph*. MIT Press, Cambridge, MA.

Kominek, J., Schultz, T., and Black, A. W. (2008). Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68.

Kwak, C., Seo, J.-H., Oh, Y., and Han, W. (2021). Efficacy of the digit-in-noise test: a systematic review and meta-analysis. *Journal of Audiology & Otology*, 26(1):10.

Leclère, T., Johannesen, P. T., Wijetillake, A., Segovia-Martínez, M., and Lopez-Poveda, E. A. (2023). A computational modelling framework for assessing information transmission with cochlear implants. *Hearing Research*, 432:108744.

Litvak, L., Delgutte, B., and Eddington, D. (2001). Auditory nerve fiber responses to electric stimulation: modulated and unmodulated pulse trains. *The Journal of the Acoustical Society of America*, 110(1):368–379.

Lyon, R. F. et al. (2011). Using a cascade of asymmetric resonators with fast-acting compression as a cochlear model for machine-hearing applications. In *Autumn meeting of the acoustical society of japan*, pages 509–512.

Mertens, G., Van de Heyning, P., Vanderveken, O., Topsakal, V., and Van Rompaey, V. (2022). The smaller the frequency-to-place mismatch the better the hearing outcomes in cochlear implant recipients? *European Archives of Oto-Rhino-Laryngology*, 279(4):1875–1883.

National Institute on Deafness and Other Communication Disorders (2024). Quick statistics about hearing. Technical report, National Institutes of Health, Washington, DC. Accessed: 2025-04-14.

Newman, R. S., Morini, G., Shroads, E., and Chatterjee, M. (2020). Toddlers' fast-mapping from noise-vocoded speech. *The Journal of the Acoustical Society of America*, 147(4):2432–2441.

Nogueira, W., Litvak, L., Edler, B., Ostermann, J., and Büchner, A. (2009). Signal processing strategies for cochlear implants using current steering. *EURASIP Journal on Advances in Signal Processing*, 2009:1–20.

Oppenheim, A. V. (1999). *Discrete-time signal processing*. Pearson Education India.

Park, J.-Y., Tsukamoto, M., Tanaka, M., and Kamitani, Y. (2023). Sound reconstruction from human brain activity via a generative model with brain-like auditory features.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251.

Penfield, W. and Jasper, H. (1954). *Epilepsy and the functional anatomy of the human brain.* Little, Brown & Co.

Perraudin, N., Balazs, P., and Søndergaard, P. L. (2013). A fast griffin-lim algorithm. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4.

Polspoel, S. (2024). *Global access to digits-in-noise hearing tests: a fully automatic test development procedure*. Phd-thesis - research and graduation internal, Vrije Universiteit Amsterdam.

Rattay, F. (1986). Analysis of models for external stimulation of axons. *IEEE Transactions on Biomedical Engineering*, 33(10):974–977.

Rattay, F., Lutter, P., and Felix, H. (2001). A model of the electrically excited human cochlear neuron: I. contribution of neural substructures to the generation and propagation of spikes. *Hearing research*, 153(1-2):43–63.

Reiss, L. A., Turner, C. W., Erenberg, S. R., and Gantz, B. J. (2007). Changes in pitch with a cochlear implant over time. *Journal for the Association for Research in Otolaryngology*, 8:241–257.

Rosen, S., Faulkner, A., and Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, 106(6):3629–3636.

Schwarz, J. R., Reid, G., and Bostock, H. (1995). Action potentials and membrane currents in the human node of ranvier. *Pflugers Arch*, 430(2):283–292.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304.

Shannon, R. V., Zeng, F.-G., and Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *The Journal of the Acoustical Society of America*, 104(4):2467–2476.

Shehabi, A., Plack, C. J., Prendergast, G., Munro, K. J., Stone, M. A., Laycock, J., AlJasser, A., and Guest, H. (2025). Online arabic and english digits-in-noise tests: Effects of test language and at-home testing. *Journal of Speech, Language, and Hearing Research*, 68(1):388–398.

Slaney, M. (1998). A matlab toolbox for auditory modeling work.

Smits, C., Theo Goverts, S., and Festen, J. M. (2013). The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *The Journal of the Acoustical Society of America*, 133(3):1693–1706.

Smits, C., Watson, C. S., Kidd, G. R., Moore, D. R., and and, S. T. G. (2016). A comparison between the dutch and american-english digits-in-noise (din) tests in normal-hearing listeners. *International Journal of Audiology*, 55(6):358–365.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

Stronks, H. C., van Deurzen, R., Jansen, P. L., Briaire, J. J., and Frijns, J. H. M. (2025). Effect of speech material and scoring method on psychometric curves for cochlear implant users and typical hearing listeners. *Ear and Hearing*, pages 10–1097.

van Gendt, M. J., Briaire, J. J., Kalkman, R. K., and Frijns, J. H. (2016). A fast, stochastic, and adaptive model of auditory nerve responses to cochlear implant stimulation. *Hearing research*, 341:130–143.

Verveen, A. and Derksen, H. (1968). Fluctuation phenomena in nerve membrane. *Proceedings of the IEEE*, 56(6):906–916.

Vroegop, J., Rodenburg-Vlot, M., Goedegebure, A., Doorduin, A., Homans, N., and van der Schroeff, M. (2021). The feasibility and reliability of a digits-in-noise test in the clinical follow-up of children with mild to profound hearing loss. *Ear and Hearing*, 42(4):973–981.

Warland, D. K., Reinagel, P., and Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, 78(5):2336–2350. PMID: 9356386.

Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (1991). Better speech recognition with cochlear implants. *Nature*, 352(6332):236–238.

Yeomans, J. S. (1979). The absolute refractory periods of self-stimulation neurons. *Physiology & Behavior*, 22(5):911–919.

Zilany, M. S. A. and Bruce, I. C. (2023). Source code for the bruce, erfani and zilany (2018) auditory nerve model (dec 2023 update). https://www.ece.mcmaster.ca/~ibruce/zbcANmodel/zbcANmodel.htm. Accessed: 2025-04-17.

Appendix

Table 2: The 15 analysis bands used by the SpecRes speech coding strategy. The lower and upper bounds of each band are listed in Hz.

Band	**Lower bound**	**Upper bound**
1	306	442
2	442	578
3	578	646
4	646	782
5	782	918
6	918	1054
7	1054	1257
8	1257	1529
9	1529	1801
10	1801	2141
11	2141	2549
12	2549	3025
13	3025	3568
14	3568	4248
15	4248	8054