From Microphone to Phoneme: An End-to-End Computational Neural Model for Predicting Speech Perception with Cochlear Implants

Tim Brochier, Josef Schlittenlacher, Iwan Roberts, Tobias Goehring, Chen Jiang, Deborah Vickers, Manohar Bance

***Abstract*— *Goal:* Advances in computational models of biological systems and artificial neural networks enable rapid virtual prototyping of neuroprostheses, accelerating innovation in the field. Here, we present an end-to-end computational model for predicting speech perception with cochlear implants (CI), the most widely-used neuroprosthesis. *Methods:* The model integrates CI signal processing, a finite element model of the electrically-stimulated cochlea, and an auditory nerve model to predict neural responses to speech stimuli. An automatic speech recognition neural network is then used to extract phoneme-level speech perception from these neural response patterns. *Results:* Compared to human CI listener data, the model predicts similar patterns of speech perception and misperception, captures between-phoneme differences in perceptibility, and replicates effects of stimulation parameters and noise on speech recognition. Information transmission analysis at different stages along the CI processing chain indicates that the bottleneck of information flow occurs at the electrode-neural interface, corroborating studies in CI listeners. *Conclusion:* An end-to-end model of CI speech perception replicated phoneme-level CI speech perception patterns, and was used to quantify information degradation through the CI processing chain. *Significance:* This type of model shows great promise for developing and optimizing new and existing neuroprostheses.**

***Index Terms*— neural prostheses, cochlear implants, computational models, automatic speech recognition, signal processing, information transmission, neural networks**

I. Introduction

The "fail fast, fail often" mentality, which celebrates risk and encourages rapid iterations of development and testing, has driven innovation in computer technology and business in the 21st century. This mindset, for the most part, has not been applied to the neuroprostheses industry, which includes technologies such as cochlear implants, bionic eyes, motor prostheses, spinal cord stimulators, deep brain stimulators, and brain-computer interfaces. In many ways, the conservative nature of the neuroprostheses industry is beneficial; the strict regulatory environment ensures a reliable standard of care, and risk and failure are to be avoided rather than celebrated when electrically stimulating neural tissue for medical applications. However, especially after a medical device has been approved for clinical use, further improvements are generally incremental. Even the slightest change to signal processing algorithms or stimulation parameters requires lengthy (and costly) clinical trials.

Recent developments in computational modelling of biological systems and deep neural networks0 may allow for rapid prototyping of neuroprostheses, aiding the selection of promising strategies prior to enormous development investments and potentially risky human trials. This modelling approach can greatly reduce costs and extend the range of possible prototypes and strategies, even to very unconventional ones. While animal studies give great insights into neural responses, the differing end organ anatomy and resultant differing electrical spread makes extrapolation to humans difficult. In this research, our primary aim is to simulate behavioral responses using modelled neural responses to electrical stimulation, incorporating realistic human anatomy. We focus on the cochlear implant (CI), the most widespread neuro-stimulation device, but the techniques employed can be extended to any neuroprosthesis.

The CI, which provides a sense of hearing to people with severe to profound hearing loss, is the most successful neural prosthesis to date, both in terms of number of users (more than 600,000 worldwide[2]) and the effectiveness of the sensory restoration. By activating electrodes implanted within the cochlea, CIs bypass damaged sensory receptors and directly stimulate auditory nerve fibers (ANFs), eliciting a sensation of sound. CI sound processors convert acoustic sounds to electrical pulse sequences, which are sent to the implant to generate neural excitation patterns that meaningfully represent the acoustic sounds. While most CI listeners can understand speech in quiet conditions, many have difficulty understanding speech in noise[3]. This difficulty arises because the process of encoding and transmitting acoustic information through CIs, and then through the electrically conductive medium of the inner ear, reduces spectral and temporal information that is essential for comprehending speech[4].

Progress on methods to improve CI speech perception is restricted by the cost, time and logistical requirements to conduct research studies with human CI listeners. The preferred method of evaluating new processing strategies or stimulation techniques for CIs is a double-blind study, where participants are tested before and after a period of training with a new strategy. This approach introduces a number of issues and challenges. First and foremost, experimenters do not want to induce any maladaptive plasticity by letting research participants use a suboptimal strategy for an extended period. Additionally, studies with CI listeners can be influenced by several factors that are unrelated to raw information transmission, such as attention, cognitive ability, rehabilitation methods, and duration of implant use, making it difficult to interpret results of a study. Bias is also an issue; research participants have usually had years of experience

TB and MB were supported by the Cambridge Hearing Trust, the Evelyn Trust, and the HB Allen Trust. IR was funded by the Rosetrees Trust Enterprise Fellowship (EF2020\100099), RNID Flexigrant (F112), and by the Evelyn Trust. TG was funded by a Career Development Award (MR/T03095X/1) from the Medical Research Council, UK (MRC). DV was funded by a MRC Senior Fellowship in Hearing (MR/S002537/1) and a National Institute Health Research Programme Grant for Applied Research (NIHR 201608). CJ was supported by the Wellcome Trust (204845/Z/16/Z). TB, IR, DV, and MB are with the University of Cambridge (UoC), Department of Clinical Neuroscience. TG is with the UoC, MRC – Cognition and Brain Sciences Unit. JS is with the University of Manchester, Division of Human Communication, Development, and Hearing. CJ was with the UoC, and is currently with Tsinghua University.

with their existing strategy, which might offset any advantage provided by the new strategy. Furthermore, it takes many months or even years for maximal speech perception to emerge, necessitating extended take-home trials.

Computational models of speech perception with CIs can potentially be used to more rapidly and objectively identify processing strategies that may improve information transmission in CIs, and to gain a better understanding of the underlying mechanisms and interactions with simulated biological parameters. To succeed, these models must make use of similar phonemic cues to CI listeners in decoding speech. The aim of this research is to develop a comprehensive and biologically-plausible computational model of CI speech perception, and to compare phoneme-level information transmission between the model and human CI listeners. We combine a finite element model (FEM) of an implanted cochlea, a computational model of the auditory nerve, and an automatic speech recognition neural network (ASR) to generate predictions of CI speech perception, and use this with the signals at different points along the biological pathway from sound to auditory nerve in order to understand where the bottlenecks are.

FEMs of the implanted cochlea are well-established, and when coupled with biophysical models of ANFs, they are a powerful tool for predicting complex interactions between anatomy, electrophysiology, and stimulation parameters that cannot be captured with simpler phenomenological models. This approach has been used to investigate current spread and current focusing [5],[6], site of action potential initiation [7],[8], effects of electrode placement [9], and effects of stimulus polarity [10]. Our aim is to extend the use of FEMs to predict neural responses to speech stimuli through a CI, and to evaluate these neural responses with an ASR.

Typical ASRs utilize Hidden Markov Models (HMM) and/or neural networks trained on time-frequency representations of speech such as spectrograms or mel-frequency cepstral coefficients (MFCCs). Computational models of the peripheral auditory system can accurately predict neural excitation patterns (neurograms) in response to acoustic stimuli [11]-[13] and to electric stimulation by CIs [7],[14]-[16]. These neurograms have been used for training ASRs to replicate human behavioral results. ASRs trained on neurograms from acoustic-hearing models have accurately predicted normal-hearing behavioral results for closed-set word recognition in noise [17],[18], pitch perception [19], and sound localization [20]. ASRs trained on CI outputs or neurograms from electric-stimulation models have predicted closed-set word recognition rates in noise for CI listeners and have provided some insights into factors that may underlie variability in CI outcomes, such as current spread, neural survival, and cognitive noise [21]-[23]. These ASR models of CI speech perception used simplified current spread models based on exponentially-decaying functions rather than an FEM, and phenomenological integrate-and-fire models [14],[16],[17] rather than biophysical models of ANFs.

While previous ASR models of CI speech perception have successfully estimated word recognition rates, it is unknown whether ASR models "perceive" speech in a similar way to CI users. If an ASR is to be useful at predicting speech outcomes for different CI processing strategies and stimulation techniques, it is crucial that it makes use of similar phonemic cues to CI listeners, which will be manifested by making similar phonemic errors and confusions to CI listeners. An advantage of modeling phoneme recognition rather than word recognition is that it is a lower-level neural function and less contaminated by upstream cortical effects that cause variability in speech perception, such as cognitive skills or short term memory. The knowledge gained from phoneme confusion matrices also gives more information about where to improve CI or signal processing, for example in a given frequency range. The ability of the model to replicate CI speech perception patterns depends upon how closely it can mimic the transmission and degradation of speech information through the CI processing chain.

Our end-to-end ASR model (CI-ASR) is the first to combine an FEM, a biophysical neural model, and a neural network ASR to predict phoneme recognition patterns in CI listeners. We construct an FEM of the implanted cochlea, and measure the voltage distribution along 1500 modelled auditory nerve fibers. We then use that voltage distribution to activate a biophysical model for each of the nerve fibers [15]. A phenomenological back-end is then applied to the biophysical nerve fiber model to incorporate temporal properties of neurons, and neurograms are generated for a large corpus of 4620 phoneme-labelled sentences (the TIMIT database [24]). An ASR consisting of two consecutive artificial neural networks is then trained with these neurograms; the first was trained to predict phonemes based on the input neurogram features, and the second was trained to adjust phoneme predictions based on contextual cues.

We train and test the CI-ASR on neurograms generated by the model, and evaluate the transmission of phonemic information using Information Transmission (IT) analysis [25]. We then compare IT results to data measured in human CI listeners, using phoneme confusion matrices from Donaldson and Kreft (2006) [26], McKay and McDermott (1993) [27] and Munson et al (2003) [28]. We use the CI-ASR to investigate the degradation of mutual information through the CI signal processing chain, in order to identify the bottleneck for information flow. To assess whether the CI-ASR model is sensitive to changes in stimulation parameters in ways similar to CI users, we virtually ran two "classic" CI behavioral experiments and compared results between the CI-ASR and CI listeners. The first measured speech recognition rates for different numbers of active electrodes [29], and the second measured speech recognition thresholds for different signal-to-noise ratios in babble noise [4]. It was hypothesized that if the CI-ASR modelled the transmission and degradation of speech information with CIs accurately, the IT results would match those of CI listeners.

II. Methods

A. Model Overview

Figure 1 shows a Block diagram of the basic signal processing to generate neurogram representations of speech. The CI processor pre-processes the speech signal by applying pre-emphasis, noise removal, and automatic gain control. These pre-processing steps were all applied using the default settings in the open-source Advanced Bionics Generalized

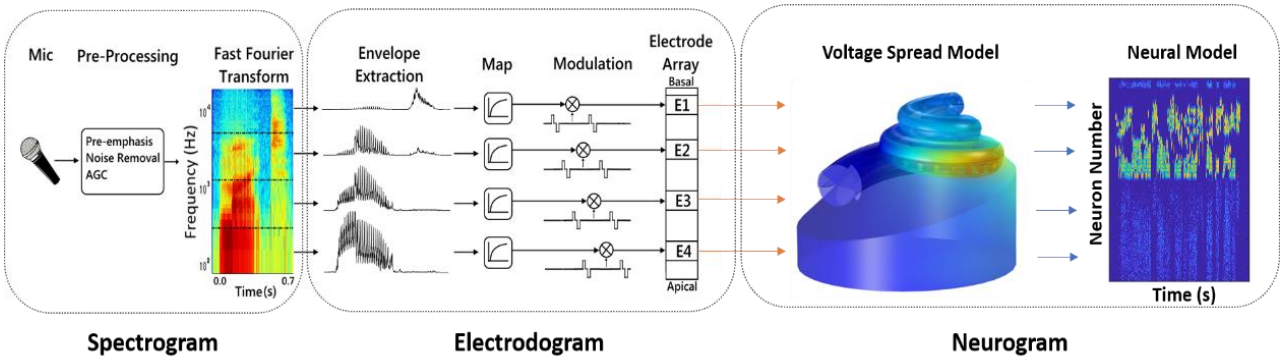

Figure 1. Block diagram demonstrating the stages of signal processing through the model.

MATLAB Toolbox (https://github.com/jabeim/GMT, demo4_procedural.m).

A spectrogram frequency decomposition is performed, and envelopes are extracted in frequency bins corresponding to each of the implanted electrodes. Note, the diagram only shows this process for four electrodes, but the CI-ASR uses a 16-electrode array. Envelopes in each channel are used to modulate interleaved biphasic pulse trains, which activate the FEM. Voltages are extracted from the FEM to activate biophysical models of neurons, whose activity over time generates the neurogram representations of speech used to train and test the ASR.

B. Finite Element Model of the Implanted Cochlea

The FEM of the implanted cochlea was implemented in COMSOL Multiphysics, and voltage spread was calculated using the electric currents interface in the AC/DC module (see Figure 2). MATLAB (version 2019b) was used in conjunction with COMSOL to automate certain elements of model generation and data extraction. A simplified cross-section of the cochlea, including the scala tympani, scala vestibuli, scala media, basilar membrane, Reisner's membrane, stria vascularis, and the osseous spiral lamina, was extruded along a parametric curve that defined the spiral of the cochlea [30]-[34]. Details of the construction of the FEM model, including the equations and conductivities used, can be found in the Supplementary Materials.

Voltages in the FEM model were used to drive the computational model of the auditory nerve. The curved plane created by the consecutive spirals within the modiolus (Figure 3) depicts the trajectory of the modelled auditory nerve fibers. Each of the consecutive spirals is 50 um apart, with descending *z*-values according to Equation 1:

$$z(n,\theta) = \frac{1}{1 + e^{-k_{\text{slope}}(n - k_{\text{angle}} * \frac{\theta}{910.3})}} \qquad (1)$$

In this equation, n represents the number of the spiral and theta represents the angle along the cochlea. The constant k_{slope} = 0.4 determines the rate of downward trajectory of neurons, and the angle dependence of that trajectory. The neural trajectory is dependent upon the radius of the cochlea at a particular point, with the nerve fibers dropping off more quickly at the apex than at the base of the cochlea. The variable k_{angle} = 0.5 controls for this angle dependence. Parameters were chosen so that the resulting neural trajectories resembled the path of neurons through the osseous spiral lamina and into and out of Rosenthal's canal. By sampling 1500 voltages along each of these 100 spirals, we generate a 100 by 1500 matrix, with each column representing the voltage along an individual auditory nerve fiber radial from the center of the main spiral.

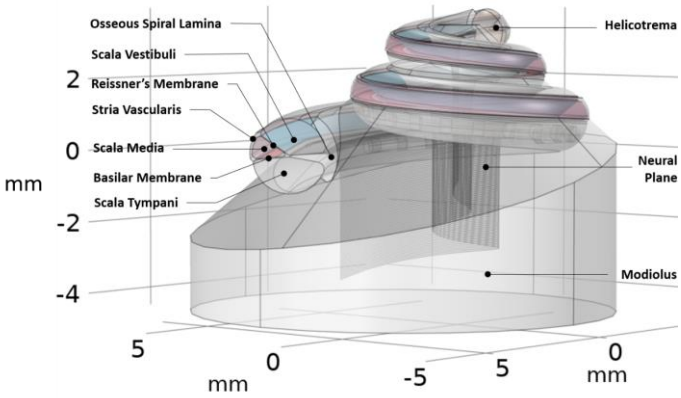

Figure 2. The parametric finite element model of the implanted cochlea, with labelled anatomical regions. The entire structure is embedded in a sphere of temporal bone with a 5 cm radius, and implanted with an electrode array.

C. Computational Model of the Auditory Nerve

The computational model of the auditory nerve was based on the biophysical multicompartmental cable model of a single auditory nerve fiber [7],[15],[16],[36]. The multicompartment cable model consists of a peripheral axon, a pre-somatic region, an unmyelinated soma (cell body), a postsomatic region, and a central axon. The axons are made up of several unmyelinated nodes of ranvier, separated by myelinated passive internodes. Each element of the cable model is characterized by a circuit with resistive and capacitive components, which is based off the foundational work by Hodgkin and Huxley (1952) [37]. The open-source Python implementation of the nerve fibers from Bachmaier et al (2019) [36] was adjusted to model a population of 1500 nerve fibers, using the morphology of the Briare and Frijns (2005) [15] nerve fiber.

To stimulate the model, the voltage at each node of Ranvier was extracted from the FEM described in the previous section. Because the FEM is purely resistive, the voltage at each node of Ranvier scales linearly with the amount of current applied to the active electrode. Hence, voltages were only extracted from the FEM at a current level of 1 mA, and these voltages were scaled to generate intracochlear voltages for a biphasic

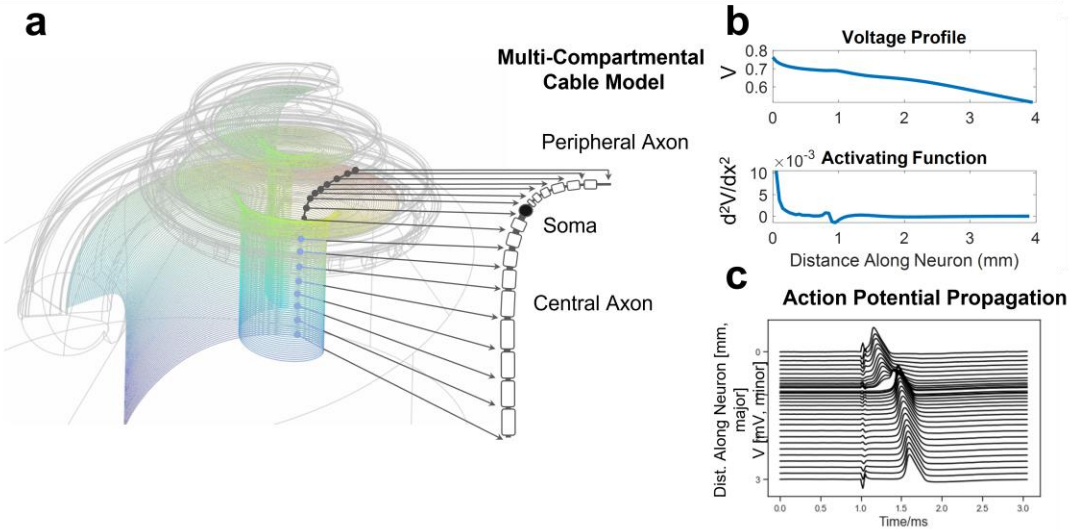

Figure 3. ***a.*** *Voltages from the finite element model are sampled along the neural trajectory, and used to activate the multi-compartmental cable model of the auditory nerve fibres. The cable model consists of a peripheral process, and pre-somatic region, a soma, a post-somatic region, and a central axon.* ***b.*** *The voltage profile along each neuron is calculated, along with the activating function (the second spatial derivative of the voltage profile). Positive values for the activating function lead to membrane depolarization.* ***c.*** *An action potential is initiated at the maximum of the activating function. In this case, the action potential is initiated at the tip of the peripheral process, and propagates past the soma and down the central axon.*

pulse (phase duration = 25 us, interphase gap = 8 us) for 300 current levels between 25 and 72 dB re 1 μA in steps of 0.16 dB.

In this way, an excitation profile was generated for every electrode of the simulated CI. These excitation profiles show which of the modelled auditory nerve fibers generate a spike as the current at the active electrode is increased from 25 dB to 72 dB re 1 μA. As the input current is increased, more auditory nerve fibers are recruited and fire.

In a binary excitation profile, each neuron fires according to a step function, where above a certain threshold the neuron always fires. To incorporate stochasticity into the firing response, these step functions were replaced with sigmoidal functions with a relative spread [38] of 0.09 ± 0.03. The relative spread is the dynamic range of a nerve fiber divided by its threshold. The dynamic range is the range of current levels over which the fiber goes from a probability of firing of 0.1 to 0.9, and the threshold is defined as the current level at which the probability of firing is 0.5.

It would be computationally intractable to calculate the cable model solution for each of the 1500 nerve fibers for the entire stimulation sequence for all the sentences in the TIMIT database. A more efficient approach is to calculate the thresholds for each neuron in response to each electrode using the cable model, and then to adjust these thresholds based on validated phenomenological models of temporal properties of auditory nerves. Not only is this approach preferrable in terms of computational efficiency, but multi-compartment cable models do not accurately predict some temporal properties of neurons such as refractoriness and adaptation [36]. For these reasons, we use the excitation profiles as an initial estimate for each neuron's activation thresholds for each electrode, and then continuously adjust those thresholds using models of refractoriness [14], adaptation [39], and temporal integration [40][41], which are described in section 2 of the Supplementary Materials. To validate the computational model of the auditory nerve, comparisons were made to electrophysiological animal data [51][52], and those comparisons are also included in the Supplementary Materials.

D. Cochlear Implant Processing Strategy

The strategy used to make comparisons in this paper is a 16 channel continuous interleaved sampling (CIS) strategy, implemented using the open-source Advanced Bionics Generalized M ATLAB Toolbox (github.com/jabeim/GMT). The pulse rate was 1250 pps per electrode, for a maximum total stimulation rate of 20000 pps across all electrode channels. Monopolar stimulation and biphasic pulses with a phase duration of 25 μs and an interphase gap of 8 μs were used. Maximum comfort levels (MCL) were set for each electrode. MCL was defined as the level at which 160 modelled neurons were activated, representing approximately 4 mm of activation along the simulated cochlea[10]. Threshold (T) levels were set to 50% of the MCL for each electrode. The parameters for the processing strategy were chosen to model speech perception patterns for an average CI user. The average stimulation rate used in [26] was 1541 pps, with a median of 866 pps, and the majority of subjects used CIS-like strategies.

E. Automatic Speech Recognizer

The ASR consisted of two sequential gated recurrent unit (GRU) artificial neural networks, both of which predict phoneme probabilities for a given time frame. We predicted the 39 standard phonemes plus the glottal stop. For the purpose of the present paper we are interested in evaluating error rates and information of the frames. Adding a HMM to transform the per-frame information to phoneme sequences yields similar phoneme error rates than similar architectures [47] of about 18%. The source code of our ASR is available at github.com/js2251/ASRbasic. A detailed description of the

ASR development can be found in the Supplementary Materials.

The first GRU network was a causal network, with an input layer, two hidden layers of 64 units each, and a dense output layer. The inputs of the causal network were the neural activation patterns for the duration of the 100 ms windows and the outputs of the causal network were the phoneme probabilities for each window. Hyperbolic tangent activation functions were used for the two hidden layers, and a softmax activation function was used for the dense output layer.

The second network was bidirectional, with an input layer, a pooling layer, two hidden layers of 128 units each, and a dense output layer. The inputs of the bidirectional network were the phoneme probabilities generated by the causal network, using a 610 ms window centered around the frame to predict, and the outputs of the bidirectional network were the phoneme probabilities of the previous, the current, and the next phoneme. Importantly, the non-causal bidirectional network allowed the ASR to update the predicted phoneme based on a wider temporal context.

The ASR was trained on neurograms generated for all 4620 sentences from the training set of the TIMIT database, with 90% of those sentences used for training and 10% of those sentences used for validation. It was then tested on neurograms for the 1680 sentences in the testing set.

The ASR was trained in 20-talker babble noise, with the SNR evenly distributed between 0 and 20 dB SNR. The ASR was then tested with noise-free speech, to replicate the studies in CI listeners that used speech in quiet for the evaluation. This procedure was used to reproduce typical acoustic environments encountered by human listeners, which often entail hearing speech in background noise.

We used this training scheme for the ASR to promote learning of noise-robust speech features that are more ecologically-valid than speech features that require quiet conditions to be salient. If the ASR were trained only with speech in quiet, it would likely learn to use very subtle cues (such as low-level modulations, high-frequency modulations, shallow modulation depths, timing differences between channels, or small spectral differences) that would not be perceptible to CI listeners.

Finally, similar to other computational models of CI speech perception [22][23], "cognitive noise" was applied to the neurograms by summing Gaussian pink noise to each channel. The pink noise was scaled to produce an SNR of 10 dB for the healthy neural condition. This SNR was inherently reduced when neural degeneration was applied in the model.

F. Applying Neural Degeneration

Four conditions were run for the CI-ASR. In the healthy condition, all 1500 neurons were active, and all neurons had an intact peripheral axon. In the peripheral degeneration condition, all 1500 neurons were active, but the peripheral axons were removed. In the other degenerated conditions, peripheral axons were removed for all nerve fibers, and then either 50% or 75% of the neurons were deactivated to span the range of neural survival measured in cadaver studies [49][50]. One advantage of testing three different neural populations is that we can test the effect of neural degeneration on speech perception while keeping all other factors constant. Another advantage is that by measuring the variance in the ASR accuracy, we can statistically compare CI user data to the ASR data using paired t-tests.

G. Analyses of Information Transmission

Phoneme confusion matrices from both the model and CI listeners were analyzed using Information Transmission (IT; [25]). When measuring IT, multiple phoneme features are assigned to each of the stimuli, and perceptual accuracy is calculated independently for each phoneme feature. The features used for consonants were voicing, place, and manner, and the features used for vowels were f1, f2,tenseness/laxness, and duration. For consonants, voicing refers to whether the vocal folds are vibrating, place refers to the location of the vocal tract restriction, and manner refers to other factors such as tongue movement, nasality, and the degree of the vocal tract restriction. For vowels, f1 and f2 correspond to the first and second formant, and are sometimes referred to as height and frontness/backness. Tenseness/laxness describes the degree of tenseness in the tongue muscles, and duration corresponds to the length of the vowel utterance. The vowel and consonant features are provided in the Supplementary Materials, along with equations for calculating IT.

IT provides a more detailed assessment of perceptual confusions than simple percent accuracy, as they give the subject credit for correctly identifying a subset of the features of a phoneme. For example, if a subject predicted the voiced alveolar sibilant "z" when the true phoneme was the unvoiced alveolar sibilant "s", the manner and place information would still have been successfully transmitted, and the voicing information would have been unsuccessfully transmitted.

III. Results

A. Consonants

Each consonant is comprised of three features: manner of articulation (determined by degree of vocal tract restriction, tongue and lip movement, and nasality), place of articulation (determined by location of vocal tract restriction), and voicing (determined by whether or not the vocal folds are active). For consonant analysis, CI-ASR results were compared to confusion matrices from Donaldson and Kreft (2006; N=20)[26]. Confusion matrices are shown for the CI-ASR and CI listeners for consonants in Figure 4, Panel a. The clusters in the confusion matrices are similar between the CI-ASR and CI listeners for consonants, with confusions tending to cluster around manner features. For example, plosives such as 'p', 't', and 'k' are often confused with one another, as are the nasals 'm' and 'n', and the fricatives 'ch', 'sh', 's' and 'z'. The mean squared error between the model and CI listeners was 0.28%. The low value of MSE was partly due to the concentration of responses along the diagonal, and the relatively sparse number of responses elsewhere. IT results are shown in Figure 4, panel b. For all measures and all features, no significant differences were found between the ASR results and the CI listener data. Similar to CI listeners, the highest Transmitted/ Input efficiency is for the manner speech feature. This can be explained as the manner feature is mostly reflected in the envelope of the signal, and because the CI exclusively encodes the envelope in each frequency band.

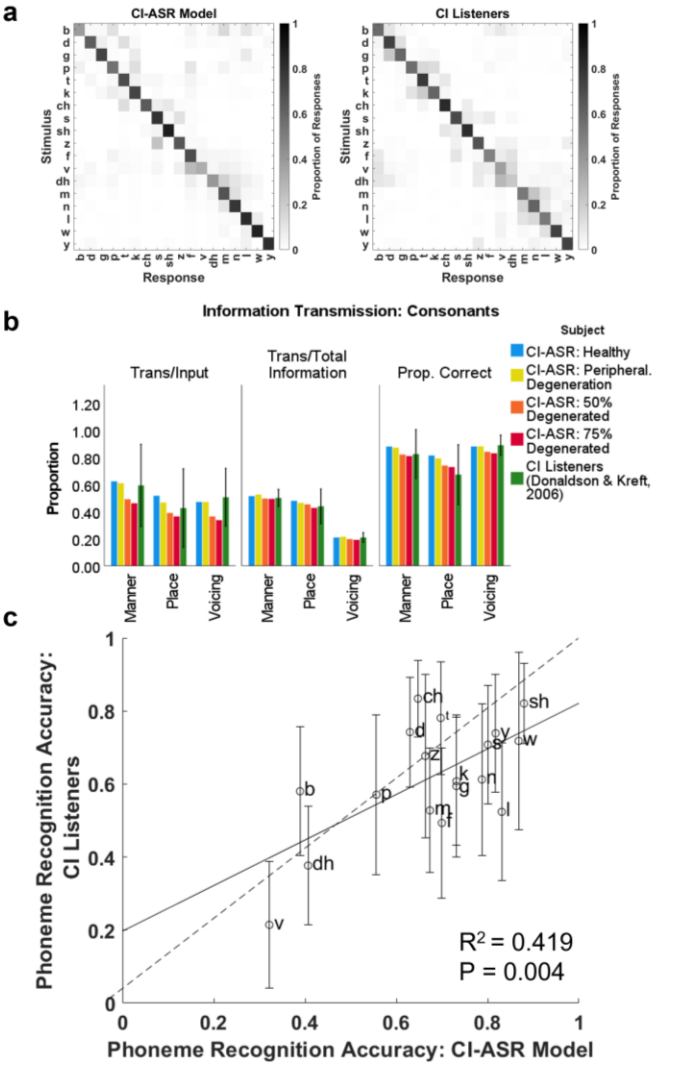

Figure 4. a. Consonant confusion matrices for the model (left) and CI listener data (right). All confusion matrices are normalised across responses (rows), and colors represent the proportion of responses. b. Comparison of consonant IT results between the model and CI listener data [26]. Each consonant is comprised of three features: manner of articulation, place of articulation, and voicing. Error bars represent the standard deviation. ***c.*** *Scatter plot for the Pearson correlation between ASR model and CI listener results for consonants. A significant correlation was found between the consonant recognition rates predicted by the model and measured in CI listeners. The dotted line represents the line of equality, where model predictions are exactly equal to CI user data, and the solid line represents the linear regression. Error bars represent the standard deviation for the CI listener data.*

The Transmitted/Total Information shows that manner and place speech cues are most important for identifying consonants for both the ASR and the CI listeners, with a significantly lower proportion of information conveyed by the voicing speech cue. Consistent with CI listeners, the model has the highest proportion correct for voicing cues and the lowest proportion correct for place cues. It is important to remember that chance performance for the binary voicing feature is 50%, while chance performance for the manner and place features, which have several options, is considerably lower. To evaluate the differences between the neural conditions simulated with the ASR (healthy, 50% and 75% degenerated), we look at the Transmitted/Input and the Proportion Correct results. The Transmitted/Total Information results are normalized to the amount of available information, so we would not expect any differences between the three neural populations. A consistent effect of neural degeneration was observed, with performance dropping as the amount of neural degeneration increased for all features.

Table I. Results from paired t-tests comparing IT results between the ASR and CI listeners for consonants.

TABLE I
PAIRED T-TEST RESULTS: INFORMATION TRANSMISSION, CONSONANTS

Feature	Transmitted/ Input	Transmitted/ Total Information	Proportion Correct
Manner	F(1,21)=0.356 p = 0.557	F(1,21)=0.159 p = 0.694	F(1,21)=0.203 p = 0.657
Place	F(1,21)=0.013 p = 0.911	F(1,21)=0.257 p = 0.617	F(1,21)=2,913 p = 0.102
Voicing	F(1,21)=2.882 p = 0.104	F(1,21)=0.387 p = 0.540	F(1,21)=2.548 p = 0.125

To further assess the CI-ASR's ability to predict phoneme perception for CI listeners, a Pearson correlation analysis was used. Model predictions were compared to the mean phoneme accuracy for CI listeners across all of the individual phonemes. Vowels and consonants were analyzed separately. For consonants, a significant correlation was found (R = 0.647, p = 0.004), suggesting that the model is able to capture between-phoneme differences in perceptibility. The scatterplot in Figure 4, panel C, visualizes which consonants are most difficult to recognize by the ASR, and how these model predictions compare to the CI user data. Both CI listeners and the model have the lowest accuracies for voiced fricatives ('v' and 'dh') and plosives ('b' and 'p'). The consonants that showed the largest difference between the ASR model and CI listeners were 'ch' and 'l'. Looking to the confusion matrices in Figure 4, the model often mistook 'ch' for other fricatives 't', 's', 'sh', and 'z', whereas CI listeners only occasionally mistook 'ch' for 's' or 't'. The model correctly identified the lateral approximant 'l' most of the time, while the CI listeners commonly confused it with nasals 'm' and 'n', or the labio-velar approximant 'w'. Paired t-tests were used to statistically assess differences between the ASR results and the CI listener data for IT. The results are summarized in Table I.

B. Vowels

For vowels, we used pooled confusion matrices from McKay and McDermott (1993; N=5)[27] and Munson et al (2003; N=30)[28]. Munson et al (2003) divided their sample into two distinct groups based on speech outcomes. The better-performing group had an average phoneme accuracy of 78.5±6.1% (n = 14), and the poorer-performing group had an average phoneme accuracy of 46.9±13.0% (n = 16). Therefore, to complete the paired t-test analyses, we considered each of the three pooled confusion matrices as one "subject" in our analysis. This approach was possible because [28] divided the subject data into better and poorer-performing groups, and together with the third dataset this

allows us to estimate the variability in the data for the statistical analysis.

Confusion matrices are shown in Figure 5, Panel a. Each vowel is comprised of four features: duration (length of pronunciation), formant 1 (frequency of the first formant), formant 2 (frequency of the second formant), and tenseness/laxness (degree of tension in the mouth and tongue muscles). For vowels, the confusion matrices show that CI listeners are more accurate than the model. The model has difficulty identifying the 'uh' vowel, and an unnatural proportion of responses are weighted towards the 'ih' vowel. The mean squared error between the model and CI listeners for vowels 0.48%. Again, the low value of MSE was due to the concentration of responses along the diagonal, and the relatively sparse number of responses elsewhere.

IT results for vowels are shown in Figure 5, Panel b. For the Trans/Input measures, the paired t-tests showed a significant difference between the CI listeners and the CI-ASR for duration cues, suggesting that CI listeners are better at using duration cues than the CI-ASR. There was no significant difference in Trans/Input measures between the CI-ASR and CI listeners for the F1, F2, or tense/lax features but this was likely due to the large variance in the human CI data. Our analysis was somewhat underpowered because we only had access to the pooled confusion matrices, rather than the individual confusion matrices used in the consonant analysis.

There was a significant difference between Trans/Total information for all features, indicating that the prioritization of phonemic cues were different for the CI-ASR compared to CI listeners. The Trans/Total information results suggest that the CI-ASR relies upon the f1 cue more than CI listeners, and that the ASR is poor at distinguishing duration and tense/lax cues. For the proportion correct measure, both the ASR and CI listeners were more accurate for f2 cues than for f1 cues. Again, CI listeners seem to be better at using duration and tense/lax cues than the ASR, and this reached statistical significance for the duration cue (p=0.034).

The different neural degeneration conditions had a consistent effect on phoneme recognition, with Transmitted/Input efficiency and Proportion Correct steadily decreasing with increased neural degeneration. A further Pearson correlation analysis was performed for vowels (Figure 5, panel c). For this comparison, we used the Munson et al (2003)[28] data.

Because we only have pooled confusion matrices for better and poorer-performing groups, the error bars here represent the range divided by two. There was not a significant correlation (R^2 =0.144, p = 0.248), indicating the ASR did not predict between-phoneme differences for vowel recognition by CI listeners. The two vowels that were most different between the model and the CI user data were 'uh' (as in book) and 'ih' (as in bit). Looking to the confusion matrices in Figure 5, the ASR model seems unable to identify the c onsonant 'uh', and responses are distributed fairly evenly across the vowels. The model also seems to default towards predicting 'ih', and regardless of the stimulus, it will often choose 'ih'. A summary of the paired t-test results for each feature and each measure is shown in Table II.

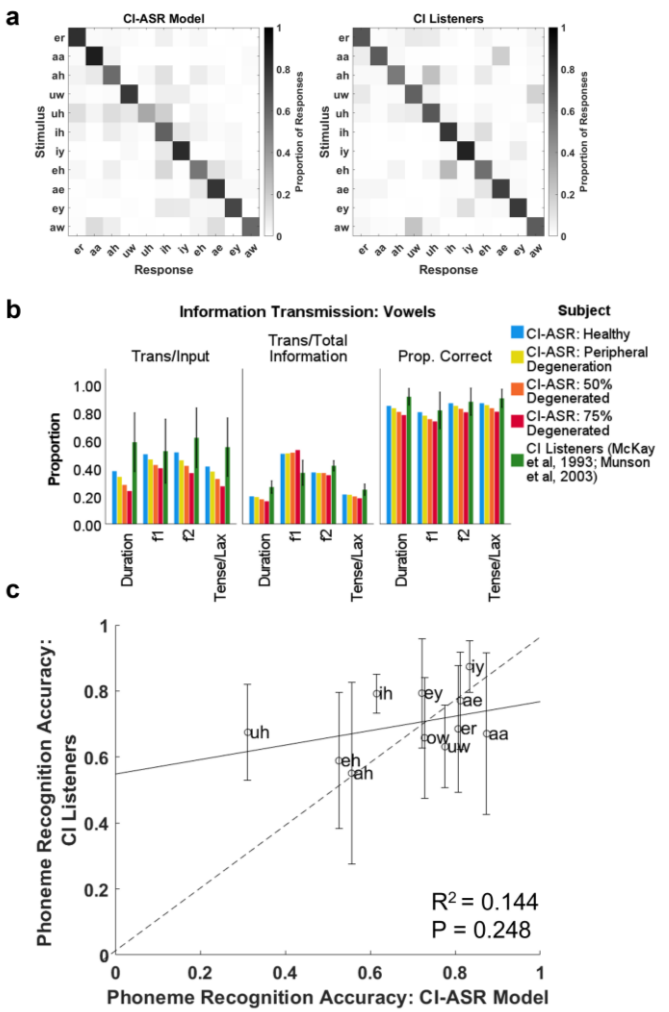

Figure 5. ***a.*** *Confusion matrices for vowels for the model (left) and CI listener data (right). All confusion matrices are normalised across responses (rows), and colours represent the proportion of responses.* ***b.*** *Comparison of vowel IT results between the model and CI listener data [27] [28]. Each vowel is comprised of four features: duration, formant 1, formant 2, and tenseness/laxness. Error bars represent the standard deviation of the pooled confusion matrices.* ***c.*** *Scatter plot for the Pearson correlation between ASR model and CI listener results for vowels. No significant correlation was found between ASR predictions and CI listener data. The dotted line represents the line of equality, where model predictions are exactly equal to CI user data, and the solid line represents the linear regression.*

Table II. Results from paired t-tests comparing IT results between the ASR and CI listeners for vowels.

TABLE II
PAIRED T-TEST RESULTS: INFORMATION TRANSMISSION, VOWELS

Feature	Transmitted/ Input	Transmitted/ Total Information	Proportion Correct
F1	F(1,4)=0.442 P=0.535	F(1,4)=11.064 **P=0.021***	F(1,4)=0.532 P=0.498
F2	F(1,4)=2.683 P=0.162	F(1,4)=8.954 **P=0.030***	F(1,4)=0.706 P=0.439
Duration	F(1,4)=6.625 **P=0.050***	F(1,4)=12.339 **P=0.017***	F(1,4)=8.392 **P=0.034***
Tense/Lax	F(1,4)=3.673 P=0.113	F(1,4)=7.641 **P=0.040***	F(1,4)=3.352 P=0.127

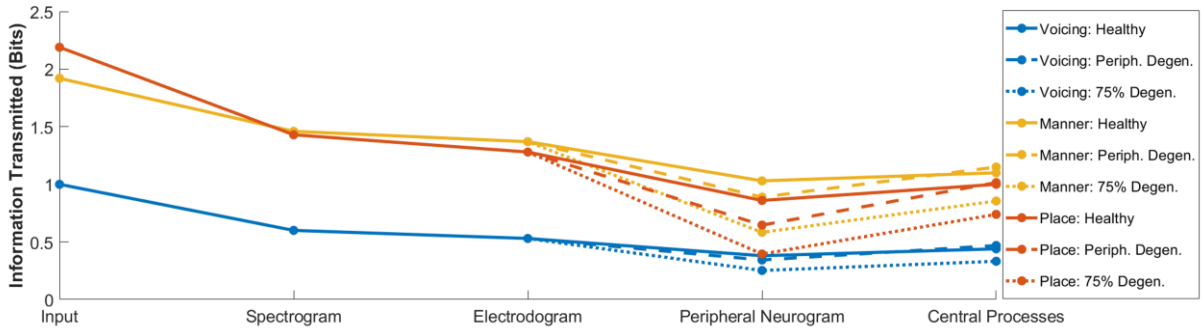

Figure 6. IT results for different points in the CI signal processing pipeline for consonants. Solid lines, dashed lines, and dotted lines represent the healthy condition, peripheral degeneration condition, and the 75% degenerated condition, respectively.

C. Probing Information Transmission through the CI Processing Chain

In the next analysis of the model, IT was probed at four different stages in the CI speech processing chain (Figure 6); the spectrogram, electrodogram, peripheral neurogram, and central neurogram. The aim of this analysis was to determine the potential patterns of IT if a CI user had access to all the information contained in the spectrogram or electrodogram. Another aim was to identify bottlenecks in IT, and to observe whether those bottlenecks differ between phonemic features. The spectrogram for normal clean speech aimed at using a maximum of the available spectral information, without using transformations like cepstral coefficients that other spectral representations like electrodograms would not use either. We extracted 198 spectral magnitude features, 33 each from six FFTs with window sizes between 4 and 128 ms.

The electrodograms were generated using the 16-channel CIS strategy described in the methods section. The ASR that was trained on spectrograms, electrodograms, and "peripheral" neurograms. Peripheral neurograms only included the causal portion of the network, because the context-based adjustment of phoneme probabilities that takes place in the non-causal network is a more central process. They were all trained in quiet and tested in quiet, so that they made use of any information that was available in the signal, thus maximizing IT potential. The "central" neurogram used both the causal and non-causal network, and included the same 10 dB SNR internal pink noise described in the methods section. This central neurogram ASR was trained in 0 – 20 dB SNR babble noise and then tested in quiet, to impose ecologically-valid limitations on the types of cues that the ASR could learn. Results of the IT analysis for consonants are presented in Figure 6, in terms of bits. Results for vowels are included in the Supplementary Materials.

IT analysis at the different stages in the CI processing pipeline are shown for the healthy condition, the peripheral degeneration condition, and the 75% degeneration condition.

For all features, the largest drop in information occurs at the electrode-neural interface, between the electrodogram and the peripheral neurogram. The magnitude of this information loss is larger for the degenerated conditions, with the 75% degeneration condition showing the largest effect. Some information is recovered with the addition of the non-causal portion of the ASR, which models context-based adjustments of phoneme predictions.

D. Replicating Studies in CI Listeners

To evaluate whether the CI-ASR and CI listeners are affected similarly by manipulations in CI processing parameters and stimuli, two common CI experiments were replicated. The first measured the effect of number of active electrodes on speech recognition in quiet, and the second measured the speech recognition threshold (SRT; SNR required to reach 50% recognition) in 20-talker babble noise.

For the number-of-channels experiment, CI-ASR results were compared to those of Schvartz-Leyzac et al (2017)[29]. MAPs were created with 8, 12, 16, and 20 channels, with channels deactivated in the same manner as [29]. The ACE n-of-m processing strategy was used with a stimulation rate of 500 pps and 8 maxima selected per frame. Results for the CI-ASR and for the CI listeners are shown in Figure 7, panel A. Three CI-ASR conditions are shown; in the Train-Xch/Test-Xch, the CI-ASR was trained and tested on the same number of channels, while in the in the Train-20ch/Test-Xch condition, the CI-ASR was trained on 20 channels and tested on either 8, 12, 16, or 20 channels. The mean between these two conditions is also shown. For both the CI-ASR and the CI Listeners, recognition rates increase with increasing number of channels. The highest increase is between 8 and 12 channels, and then recognition rates begin to level off from 16 to 20 channels. Recognition rates for the CI-ASR fall within the range of recognition rates in CI listeners for every condition.

For the SRT experiment, the CI-ASR that was trained in mixed noise was tested in 20-talker babble noise with a range of SNRs. Phoneme recognition rate as a function of SNR is shown in Figure 7, panel B, for the CI-ASR and CI listener data [4]. The SRT was 6.3 dB, which is within 1 standard deviation of the CI user average in several CI studies. For example, Friesen et al (2001)[4] measured an SRT of 6.47 ± 0.53 dB in speech-shaped noise, and Goehring et al (2017)[53] measured an SRT of 6.70 ± 0.90 dB in 20-talker babble noise.

IV. Discussion

Phoneme-level information transmission (IT) was compared between CI listeners and a neurogram-trained CI-ASR.

A. Consonant Information Transmission

For consonants, the CI-ASR had similar patterns of phoneme recognition to CI listeners. Manner cues were most

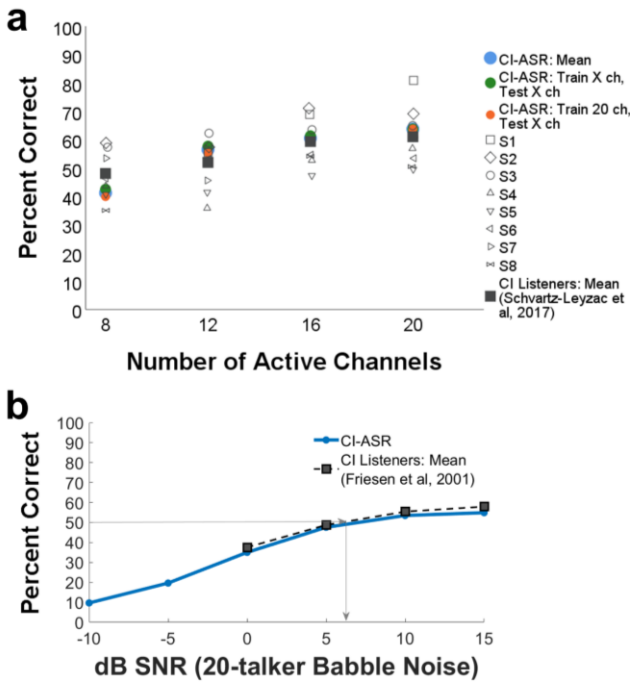

Figure 7. a. CI-ASR recognition rate as a function of number of active channels, compared to CI listener data [35]. b. CI-ASR phoneme recognition rate as a function of SNR for 20-talker babble noise, compared to CI listener data [6]. The grey line shows the speech recognition threshold (SRT) for the CI-ASR, where the recognition rate is 50%.

effectively transmitted for both CI listeners and the model, as reflected in the IT analysis. This result was expected, considering that the manner of articulation primarily affects the envelope of a phoneme [54], and that CIs encode the envelope in each stimulation channel. For the CI-ASR model and the CI user data, place cues had the lowest proportion correct. This result is consistent with several other studies [32]-[34], [55]-[56] reporting higher accuracy for manner and voicing features than for place features in CI users.

In the consonant IT analysis, no significant differences were found between the CI-ASR and CI listener data for any of the measures or features. The significant correlation ($R^2 = 0.419$, $p = 0.004$) between ASR model accuracies and mean CI listener accuracies for consonants suggests that the ASR model successfully predicted between-phoneme differences in perceptibility of human CI listeners. However, 58% of the variance in perceptibility between phonemes is still unexplained by the model, which could be due to a wide range of factors that our single average model does not account for. Now that a basic model validation has been accomplished, an improved approach would be to run a detailed model on each CI listener's specific processing strategy. In the future, we envision that the FEM could be parametrized for different CI listeners based on clinical CT scans, to more accurately estimate current spread within their individual cochleae.

Despite a higher proportion of voicing information than place information being transmitted, CI listeners do not necessarily find the voicing feature more useful than the place feature for consonant recognition. On the contrary, the Transmitted/Total Information results show that the place feature transmits more information for CI listeners than voicing features, for both the model and the CI listeners.

B. Vowel Information Transmission

The significant differences across features in Transmitted/Total Information indicate that the model prioritizes vowel features differently than CI users. For the duration feature, CI-ASR performance was below the average CI listener for all IT measurements, and the lack of significant differences for the other features was somewhat due to the large variance in the CI listener data.

The relatively poor transmission of duration cues in the ASR compared to CI listeners was expected, because the ASR used fixed-length time windows for feature extraction and phoneme identification and this quantization may have distorted the duration cues. Another factor that may have influenced the relatively poor transmission of the duration feature was that for CI listeners, the consonant context of the vowels was consistent throughout the experiments. The consonant context is known to have a large effect on vowel duration [57]. For example, Jones (1956) [58] reported that the 'ee' vowel in the word "see" has a duration of 0.317 seconds, while the same 'ee' vowel in the word "seating" has a duration of 0.087 seconds. In the TIMIT database, the vowels were presented in a wide range of consonant contexts, and the duration cue was not as reliable for identifying vowel sounds.

The CI-ASR was trained on a wide array of dialects in the United States from over one hundred speakers. As vowels are the primary phonetic feature that changes with different dialects, it is likely that the CI-ASR learned to generalize the formant cues. Combined with the limited spectral resolution of the CI, this generalized formant representation may explain some of the discrepancy between vowel IT for the model and for CI listeners. In the CI user experiments which the ASR was compared to [27][28], there were only 1-3 speakers, so the CI listeners may have been able to learn the specific vowel features for each speaker.

To test the hypothesis that generalizing phonetic cues across dialect regions would affect ASR performance, the CI-ASR (healthy neural condition) was trained on the dialect region from the Northern United States, and tested on the dialect region from the Southern United States. Percent accuracy lowered 4.9% for vowels and 4.6% for consonants compared to a CI-ASR trained and tested on the Southern United States dialect region. Information transmission reduced by 0.2 bits for vowels (15% of Transmitted/Total) and 0.17 bits for consonants (10% of Transmitted/Total). This result suggests that consonant and vowel perception are both affected by training from different dialect regions, with a slightly stronger effect on vowels.

C. Probing Information Transmission through the CI Processing Chain

The analysis of IT through different processing stages of the CI demonstrates that the bottleneck of information flow in terms of phonemes occurs between the electrodogram and the peripheral neurogram. This result corroborates many studies that have shown the relationship between the electrode-neural interface, as estimated by psychophysics or electrophysiology, and speech perception in CI listeners [59]-[63].

During the conversion from spectrogram to electrodogram, there is a small reduction in IT of about 5-7% for consonants and 2-3% for vowels, across features. This drop in information

is likely due to the reduction in frequency range and resolution that occurs when the 198 frequency bins of the spectrogram are reduced to envelopes in 16 CI channels. A larger reduction in IT of 14-16% for consonants and 7-10% for vowels occurs when the peripheral neurogram is generated from the electrodogram. This reduction is due to the spread of excitation within cochlear fluids, which disrupts channel independence. For the degenerated neural conditions the information loss up to the electrodogram is identical, and more information is lost at the electrode-neural interface, with the most information lost for the 75% degenerated condition.

Some of the lost information is restored by the central neurogram, which adjusts phoneme predictions based on context cues. The context cues are particularly important for vowel recognition, which may be due to the long window sizes in the non-causal network (610 ms) compared to those used in the causal network (100 ms). Vowel durations can be up to 350 ms, so the long window size in the non-causal network may have helped to identify longer duration vowels.

The place feature has the most information loss among consonant features, with up to 1.33 bits lost at the peripheral neurogram. The place of articulation is typically reflected in the temporal fine structure of a phoneme. For example, formant transitions help to differentiate between the bilabial nasal ‘m’ and the alveolar nasal ‘n’, and formant transitions into and out of plosives help to differentiate between bilabial ‘b’, alveolar ’d’, and velar ‘g’. These formant transitions are not encoded well in cochlear implants, particularly for higher formants [64]. The shifting frequency cue might be contained in a single frequency band, and even if the frequency shift is large enough to affect several adjacent electrodes, the targeted neural population is still similar due to current spread, especially at current levels near C-level.

For vowels, the most information loss occurred for the F1 feature, with up to 1.20 bits lost at the peripheral neurogram. The fact that the ASR was trained across several dialects may have forced it to generalize formant features. Combined with the current spread in the cochlea, a degradation in F1 transmission would be expected. Additionally, the relatively small range of F1 cues (250 – 900 Hz) compared to F2 (600 – 2400 Hz) cues might explain why more information was lost for the F1 cue than the F2 cue. For the AB default frequency allocation table, there are 3 frequency bins in the range of F1 cues, and 7 frequency bins in the range of F2 cues.

For consonants and vowels, there is an initial large decline in information simply by converting the input to the spectrogram and using the ASR. The input information is calculated assuming 100% recognition, which is not achievable by any ASR system. The spectrogram-trained ASR sets the upper limit for IT in our study. While the spectrogram is not the optimal input feature for ASRs, it is the same transform used in CI devices. From the ASR literature, other features such as first and second order delta features, linear prediction coefficients (LPC), mel-frequency cepstral coefficients (MFCCs), and the discrete wavelet transform (DWT) have all been shown to improve ASR performance. Some publications have suggested that the use of alternative time-frequency transforms at the front end of CIs may improve speech recognition [65].

Probing IT at different points in the CI signal processing chain also demonstrates that the complexity of our end-to-end model was necessary to accurately replicate patterns of CI speech perception. If a simple spectrogram or electrodogram-trained ASR were used, the model predictions of IT would have been significantly different than those of CI listeners.

D. Replicating Studies in CI Listeners

Two common CI studies were replicated by the CI-ASR, demonstrating that the CI-ASR can predict effects of modifications to stimulation parameters and input stimuli on speech recognition. This model validation was important, because the primary aim of the model is to provide a platform for the evaluation of new processing strategies and stimulation techniques. Because CI-ASR had a similar speech recognition threshold in noise to CI listeners, we expect that the model will be useful for trialing noise-removal techniques in CIs. Neural network based noise removal techniques [59] will become widespread as processors become capable of handling their computational demands. There are endless choices for the parametrization and optimization of noise removal techniques, from the type of training data to the neural network architecture. The CI-ASR may help to identify the most successful strategies prior to studies in CI-listeners and to fine-tune the parameters.

The CI-ASR was also affected similarly by the number of active electrodes, indicating that the model is replicating some of the complex interactions between current spread, neural dynamics, and speech information transmission. While the number-of-channels experiment is only one of the basic ways to alter a stimulation strategy, it is a critical benchmark before evaluating more advanced processing strategies. For example, psychoacoustic-ACE (PACE)[66], spatial-ACE (SPACE)[67], and the temporal integration processing strategy (TIPS)[68] all attempt to minimize channel interaction by removing pulses that would be spatially and/or temporally masked by previous pulses. The combined results of the IT analysis and the number-of-channels experiment suggest that the CI-ASR is accurately modelling channel interactions, which would be critical to evaluate and optimize strategies like these.

E. Limitations

There are some limitations to the CI-ASR that will be improved upon in future versions. Compared to other cochlear FEMs that use sectioned microCT scans [8][10], the parametric FEM used in this study is a simpler reconstruction that conveys the 3D spiraling and tapering structure of the cochlea. The parametric FEM replicates transimpedance matrices from CI listeners, but that only validates the current pathways within the scala tympani. There is less basal shunting in the FEM compared to CI users, and the voltage tends to be slightly higher in the FEM than in the CI listener data. Electrical measurements within the actual modiolus will be necessary to validate other important current pathways due to electrical stimulation.

While microCT reconstructions may be the preferred approach to capture fine details of the cochlea, the parametric approach could prove advantageous when trying to extract important parameters from low-resolution clinical CT scans from individual patients. For example, one could generate an

approximation for the cochlear shape using CT reconstruction [69], and extract the equation for the cochlear spiral and taper in order to generate patient-specific parametric cochlear models, or directly import microCT reconstructions of all the substructures into the model.

The conductivities of the materials in our model were purely resistive. This would not be precise compared to the real-world situation, where the presence of neurons introduces highly capacitive components, because the response of neurons with influx and outflux of ions is a slow process compared to electron movement. Therefore, for high-frequency stimulation pulses, the impedance of some materials ought to be low at the transitions between phases (such as in biphasic pulses). For future development, it would be interesting to include capacitive elements in the model to reveal the differences caused by high-frequency components. We note that other state-of-the-art FEMs in the cochlear implant literature [8][10] also use purely resistive models.

Another limitation is that the model disregards the effects of facilitation, where closely-spaced subthreshold pulses can initiate an action potential, or accommodation, where neurons lose sensitivity with prolonged stimulation [70].

The current version of the model is limited to assessing the transmission of atonal, phoneme-based speech segments through a CI. Perception of tonal language, emotional prosody, and music is known to be a challenge for CI listeners. With a tone-labelled dataset like TIMIT, this model could test processing strategies that encode pitch information.

F. Advantages and Applications

The CI-ASR is not limited by the same physical, financial, and practical factors that constrain behavioral clinical studies in CI listeners. The effect of experimental electrode arrays, site selection strategies, focused stimulation modes, and various pulse parameters could all be evaluated while providing precise control over factors such as neural degeneration, fibrosis, and electrode placement. Analyses using the model will also eliminate variance due to factors such as attention, cognitive ability, and duration of implant use. This level of control is especially important during the prototyping stage, where it would be impractical to use human trials to determine every single parameter in a processing strategy. Overall, we envision the model being used in conjunction with traditional listening studies in CI listeners, to accelerate the development of new CI strategies.

The approach used by the CI-ASR can be generalized to other neural prostheses, such as brain-computer interfaces, spinal cord stimulators, and visual, motor, and vestibular prostheses. The morphology of the cable model and the FEM would need to be adjusted according to the physiology of the intended stimulation targets. Neural excitation patterns in response to electrical stimulation could then be used to train a classification neural network.

V. Conclusions

An end-to-end model of CI speech perception, consisting of an FEM, a computational model of the auditory nerve, and an ASR, was validated and used to predict phoneme-level information transmission in CI listeners. For consonant features, no significant differences were found between the model and CI listener data. A significant correlation was found between model predictions and CI user data for consonant recognition accuracies, but not for vowel recognition accuracies. The model replicated effects of stimulation parameters and noise on speech recognition. Information transmission was also probed at different stages in the model pipeline, suggesting that the bottlenecks of information flow are the spectrogram transformation and the electrode-neural interface. This type of model will be useful for developing, optimizing, and predicting the efficacy of new CI processing strategies.

References

[1] Y. LeCun, *et al*, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] "Cochlear Implants," *NIDCD*. Available: https://www.nidcd.nih.gov/health/cochlear-implants (accessed Feb. 28, 2022).

[3] P. Blamey *et al*., "Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: an update with 2251 patients," *Audiol. Neurootol.*, vol. 18, no. 1, pp. 36–47, 2013..

[4] L. M. Friesen, *et al*, "Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants," *JASA*, vol. 110, no. 2, pp. 1150–1163, Aug. 2001.

[5] J. H. Frijns *et al*, "Potential distributions and neural excitation patterns in a rotationally symmetric model of the electrically stimulated cochlea," *Hear. Res.*, vol. 87, no. 1–2, pp. 170–186, Jul. 1995.

[6] R. K. Kalkman, J. J. Briaire, and J. H. M. Frijns, "Current focussing in cochlear implants: an analysis of neural recruitment in a computational model," *Hear. Res.*, vol. 322, pp. 89–98, Apr. 2015.

[7] F. Rattay *et al*, "A model of the electrically excited human cochlear neuron. I. Contribution of neural substructures to the generation and propagation of spikes," *Hear. Res.*, vol. 153, no. 1–2, pp. 43–63, Mar. 2001.

[8] T. Potrusil *et al*., "Finite element analysis and three-dimensional reconstruction of tonotopically aligned human auditory fiber pathways: a computational environment for modeling electrical stimulation by a cochlear implant based on micro-CT," *Hear. Res.*, vol. 393, p. 108001, 2020.

[9] T. Hanekom, "Modelling of the electrode-auditory nerve fibre interface in cochlear prostheses," 2001, [Online]. Available: https://repository.up.ac.za/handle/2263/27742 (Accessed Feb. 28, 2022)

[10] R. K. Kalkman *et al*, "The relation between polarity sensitivity and neural degeneration in a computational model of cochlear implant stimulation," *Hear. Res.*, vol. 415, p. 108413, Mar. 2022.

[11] M. S. A. Zilany, I. C. Bruce, and L. H. Carney, "Updated parameters and expanded simulation options for a model of the auditory periphery," *JASA*, vol. 135, no. 1, pp. 283–286, Jan. 2014..

[12] S. Verhulst, A. Altoè, and V. Vasilkov, "Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss," *Hear. Res.*, vol. 360, pp. 55–75, Mar. 2018..

[13] D. Baby, A. Van Den Broucke, and S. Verhulst, "A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications," *Nat Mach Intell*, vol. 3, no. 2, pp. 134–143, Feb. 2021.

[14] I. C. Bruce, M. W. White, L. S. Irlicht, S. J. O'Leary, and G. M. Clark, "The effects of stochastic neural activity in a model predicting intensity perception with cochlear implants: low-rate stimulation," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 12, pp. 1393–1404, Dec. 1999.

[15] J. J. Briaire and J. H. M. Frijns, "The consequences of neural degeneration regarding optimal cochlear implant position in scala tympani: a model approach," *Hear. Res.*, vol. 214, no. 1–2, pp. 17–27, Apr. 2006.

[16] J. E. Smit *et al*, "Threshold predictions of different pulse shapes using a human auditory nerve fibre model containing persistent sodium and slow potassium currents," *Hear. Res.*, vol. 269, no. 1–2, pp. 12–22, Oct. 2010.

[17] A. J. E. Kell *et al*, "A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy," *Neuron*, vol. 98, no. 3, pp. 630–644.e16, May 2018.

[18] S. Haro *et al* "Deep Neural Network Model of Hearing-Impaired Speech-in-Noise Perception," *Front. Neurosci.*, vol. 14, p. 588448, Dec. 2020.

[19] M. R. Saddler, R. Gonzalez, and J. H. McDermott, "Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception," *Nat. Commun.*, vol. 12, no. 1, p. 7278, Dec. 2021.

[20] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nat Hum Behav*, vol. 6, no. 1, pp. 111–133, Jan. 2022.

[21] W. Nogueira, T. Harczos, B. Edler, J. Ostermann, and A. Büchner, "Automatic speech recognition with a cochlear implant front-end," 2007. [Online]. Available: https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2007/i07_2537.pdf
[22] S. Fredelake and V. Hohmann, "Factors affecting predicted speech intelligibility with cochlear implants in an auditory model for electrical stimulation," *Hear. Res.*, vol. 287, no. 1–2, pp. 76–90, May 2012.
[23] T. Jürgens *et al* "The effects of electrical field spatial spread and some cognitive factors on speech-in-noise performance of individual cochlear implant users-A computer model study," *PLoS One*, vol. 13, no. 4, p. e0193842, Apr. 2018.
[24] V. Zue *et al* "Speech database development at MIT: Timit and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, Aug. 1990.
[25] G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," *JASA*, vol.27, no.2, pp.338–352, Mar. 1955.
[26] G. S. Donaldson and H. A. Kreft, "Effects of vowel context on the recognition of initial and medial consonants by cochlear implant users," *Ear Hear.*, vol. 27, no. 6, pp. 658–677, Dec. 2006.
[27] C. M. McKay and H. J. McDermott, "Perceptual performance of subjects with cochlear implants using the Spectral Maxima Sound Processor (SMSP) and the Mini Speech Processor (MSP)," *Ear Hear.*, vol. 14, no. 5, pp. 350–367, Oct. 1993.
[28] B. Munson *et al* "Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability," *JASA*, vol. 113, no. 2, pp. 925–935, Feb. 2003.
[29] K. C. Schvartz-Leyzac, T. A. Zwolan, and B. E. Pfingst, "Effects of electrode deactivation on speech recognition in multichannel cochlear implant recipients," *Cochlear Implants Int.*, vol. 18, no. 6, pp. 324–334, Nov. 2017.
[30] L. T. Cohen, J. Xu, S. A. Xu, and G. M. Clark, "Improved and simplified methods for specifying positions of the electrode bands of a cochlear implant array," *Am. J. Otol.*, vol. 17, no. 6, pp. 859–865, Nov. 1996.
[31] J. Wysocki, "Dimensions of the human vestibular and tympanic scalae," *Hear. Res.*, vol. 135, no. 1–2, pp. 39–46, Sep. 1999.
[32] S. K. Yoo *et al* "Three-dimensional geometric modeling of the cochlea using helico-spiral approximation," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 10, pp. 1392–1402, Oct. 2000.
[33] J. R. Clark *et al* "A Scalable Model for Human Scala-Tympani Phantoms," *J. Med. Device.*, vol. 5, no. 1, Jan. 2011, doi: 10.1115/1.4002932.
[34] K. Dang, "Electrical conduction models for cochlear implant stimulation," Université Côte d'Azur, 2017. Accessed: Feb. 28, 2022. [Online]. Available: https://hal.inria.fr/tel-01562277/
[35] C. Garcia *et al.*, "The Panoramic ECAP Method: Estimating Patient-Specific Patterns of Current Spread and Neural Health in Cochlear Implant Users," *J. Assoc. Res. Otolaryngol.*, vol. 22, no. 5, pp. 567–589, Oct. 2021.
[36] R. Bachmaier, *et al* "Comparison of Multi-Compartment Cable Models of Human Auditory Nerve Fibers," *Front. Neurosci.*, vol.13, p.1173,Nov 2019
[37] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, no. 4, pp. 500–544, Aug. 1952.
[38] A. A. Verveen, "Fluctuation in Excitability: Research Report on Signal Transmission in Nerve Fibers," 1961.
[39] M. J. van Gendt, *et al* "Short and long-term adaptation in the auditory nerve stimulated with high-rate electrical pulse trains are better described by a power law," *Hear. Res.*, vol. 398, p. 108090, Dec. 2020.
[40] A. J. Oxenham, "Forward masking: adaptation or integration?," *JASA*, vol. 109, no. 2, pp. 732–741, Feb. 2001.
[41] C. M. McKay *et al* "Temporal Processing in the Auditory System," *J. Assoc. Res. Otolaryngol.*, vol. 14, no. 1, pp. 103–124, Feb. 2013.
[42] C. M. McKay and H. J. McDermott, "Loudness perception with pulsatile electrical stimulation: the effect of interpulse intervals," *JASA*, vol. 104, no. 2 Pt 1, pp. 1061–1074, Aug. 1998.
[43] C. M. McKay *et al* "Loudness summation for pulsatile electrical stimulation of the cochlea: effects of rate, electrode separation, level, and mode of stimulation," *JASA*, vol. 110, pp. 1514–1524, Sep. 2001.
[44] H. J. McDermott, *et al*, "Application of loudness models to sound processing for cochlear implants," *JASA*, vol. 114, no. 4 Pt 1, pp. 2190–2197, Oct. 2003.
[45] T. Francart, *et al*, "Loudness of time-varying stimuli with electric stimulation," *JASA*, vol. 135, no. 6, pp. 3513–3519, Jun. 2014.
[46] N. Zhou, *et al*, "Integration of Pulse Trains in Humans and Guinea Pigs with Cochlear Implants," *J. Assoc. Res. Otolaryngol.*, vol. 16, no. 4, pp. 523–534, Aug. 2015.
[47] M. Ravanelli *et al* "The Pytorch-kaldi Speech Recognition Toolkit," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6465–6469.
[48] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12449–12460, 2020.
[49] J. B. Nadol Jr, *et al* "Survival of spiral ganglion cells in profound sensorineural hearing loss: implications for cochlear implantation," *Ann. Otol. Rhinol. Laryngol.*, vol. 98, no. 6, pp. 411–416, Jun. 1989.
[50] A. M. Khan, *et al*, "Is word recognition correlated with the number of surviving spiral ganglion cells and electrode insertion depth in human subjects with cochlear implants?," *Laryngosc.*, vol.115, no.4, pp.672–677, Apr. 2005.
[51] E. Javel *et al* "Responses of Cat Auditory Nerve Fibers to Biphasic Electrical Current Pulses," *Ann. Otol. Rhinol. Laryngol.*, vol. 96, no. 1_suppl, pp. 26–30, Jan. 1987.
[52] C. A. Miller *et al* "Changes across time in the temporal responses of auditory nerve fibers stimulated by electric pulse trains," *J. Assoc. Res. Otolaryngol.*, vol. 9, no. 1, pp. 122–137, Mar. 2008.
[53] T. Goehring, *et al* "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," *Hear. Res.*, vol. 344, pp. 183–194, Feb. 2017.
[54] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 336, no. 1278, pp. 367–373, Jun. 1992.
[55] M. W. Skinner, *et al* "Identification of speech by cochlear implant recipients with the multipeak (MPEAK) and spectral peak (SPEAK) speech coding strategies" *Ear Hear.*, vol. 20, no.6, pp.443–460, Dec. 1999.
[56] A. van Wieringen and J. Wouters, "Natural vowel and consonant recognition by Laura cochlear implantees," *Ear Hear.*, vol. 20, no. 2, pp. 89–103, Apr. 1999.
[57] V. Fridland, T. Kendall, and C. Farrington, "Durational and spectral differences in American English vowels: dialect variation within and across regions," *JASA*, vol. 136, no. 1, pp. 341–349, Jul. 2014.
[58] D. Jones, "An outline of English phonetics," *UK: Heffer*, 1956.
[59] J. A. Bierer, S. M. Bierer, and J. C. Middlebrooks, "Partial tripolar cochlear implant stimulation: Spread of excitation and forward masking in the inferior colliculus," *Hear. Res.*, vol. 270, no. 1–2, pp. 134–142, Dec. 2010.
[60] J. H. Goldwyn, S. M. Bierer, and J. A. Bierer, "Modeling the electrode–neuron interface of cochlear implants: Effects of neural survival, electrode placement, and the partial tripolar configuration," *Hear. Res.*, vol. 268, no. 1, pp. 93–104, Sep. 2010.
[61] C. J. Long *et al.*, "Examining the electro-neural interface of cochlear implant users using psychophysics, CT scans, and speech understanding," *J. Assoc. Res. Otolaryngol.*, vol. 15, no. 2, pp. 293–304, Apr. 2014.
[62] L. DeVries, R. Scheperle, and J. A. Bierer, "Assessing the Electrode-Neuron Interface with the Electrically Evoked Compound Action Potential, Electrode Position, and Behavioral Thresholds," *J. Assoc. Res. Otolaryngol.*, vol. 17, no. 3, pp. 237–252, Jun. 2016.
[63] K. N. Jahn and J. G. Arenberg, "Electrophysiological Estimates of the Electrode–Neuron Interface Differ Between Younger and Older Listeners With Cochlear Implants," *Ear Hear.*, vol. 41, no. 4, p. 948, 2020.
[64] M. B. Winn, *et al* "Assessment of Spectral and Temporal Resolution in Cochlear Implant Users Using Psychoacoustic Discrimination and Speech Cue Categorization," *Ear Hear.*, vol. 37, no. 6, pp. e377–e390, 2016.
[65] J. Yao and Y.-T. Zhang, "The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 11, pp. 1299–1309, Nov. 2002.
[66] W. Nogueira, *et al*, "A psychoacoustic 'NofM'-type speech coding strategy for cochlear implants," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 18, Dec. 2005
[67] F. Bolner, *et al*, "Precompensating for spread of excitation in a cochlear implant coding strategy," *Hear. Res.*, vol. 395, p. 107977, Sep. 2020.
[68] W. Lamping, *et al*, "The effect of a coding strategy that removes temporally masked pulses on speech perception by cochlear implant users," *Hear. Res.*, vol. 391, p. 107969, Jun. 2020.
[69] A. H. Gee, Y. Zhao, G. M. Treece, and M. L. Bance, "Practicable assessment of cochlear size and shape from clinical CT images," *Sci. Rep.*, vol. 11, no. 1, p. 3448, Feb. 2021.
[70] J. Boulet, M. White, and I. C. Bruce, "Temporal Considerations for Stimulating Spiral Ganglion Neurons with Cochlear Implants," *J. Assoc. Res. Otolaryngol.*, vol. 17, no. 1, pp. 1–17, Feb. 2016.