Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 175243, 9 pages
doi:10.1155/2009/175243

Research Article

Prediction of Speech Recognition in Cochlear Implant Users by Adapting Auditory Models to Psychophysical Data

Svante Stadler and Arne Leijon (EURASIP Member)

Sound and Image Processing Lab, KTH, 10044 Stockholm, Sweden

Correspondence should be addressed to Svante Stadler, svante.stadler@ee.kth.se

Received 15 December 2008; Revised 6 May 2009; Accepted 18 June 2009

Recommended by Hugo Fastl

Users of cochlear implants (CIs) vary widely in their ability to recognize speech in noisy conditions. There are many factors that may influence their performance. We have investigated to what degree it can be explained by the users' ability to discriminate spectral shapes. A speech recognition task has been simulated using both a simple and a complex models of CI hearing. The models were individualized by adapting their parameters to fit the results of a spectral discrimination test. The predicted speech recognition performance was compared to experimental results, and they were significantly correlated. The presented framework may be used to simulate the effects of changing the CI encoding strategy.

1. Introduction

Early cochlear implants, using only a single channel, were useful for identifying environmental sounds and improving lip reading performance. However, speech recognition with such implants was very limited. Since then, the number of channels in implants has steadily increased as technology has matured, and modern implants make use of up to 22 separate channels (the Cochlear Nucleus implant), or even up to 120 "virtual" channels (Advanced Bionics HiRes120) [1]. The theoretical basis behind this development has been that each channel stimulates mainly in a local region of the cochlea, which along with the tonotopic organization of the auditory nerve corresponds to a frequency specific sensation [2]. According to this principle, each channel can be used to encode the signal components in the corresponding frequency band. However, it has been shown repeatedly that the spectral resolution ability of CI users is limited to about 4–8 "effective" channels, even if the actual number of channels is much larger [3–5]. In an experiment using normal hearing listeners and speech signals with modified spectral content, Fu et al. [6] showed that increasing the number of effective channels increases speech recognition in noise, up to and beyond 16 channels. Therefore, it seems logical to conclude that spectral resolution is a limiting factor for CI users' speech recognition. However, it is also clear that cognitive factors such as short-term memory and lexical knowledge have an impact [7], as well as patchy or incomplete nerve survival.

In this paper, we examine the role of spectral resolution on speech recognition ability. To this end, a sentence recognition task has been simulated to predict the signal-to-noise ratio at which words are barely recognized. The simulation includes a model of CI hearing, which has been individualized using data from psychophysical experiments. This approach is not entirely novel; several earlier attempts have been made to predict speech recognition performance for hearing impaired listeners, either using heuristics [8, 9], or auditory models [10–12]. All of these studies use the audiogram as the main psychophysical data source to adapt the model. However, in the case of CI hearing, the audiogram is irrelevant, since sound levels can be arbitrarily matched to stimulation levels. Therefore, other experimental data must be used. In [5], the results of a spectral resolution experiment is found to have a limited but significant correlation with the speech recognition threshold (SRT) in noise. The data from that study were used to evaluate the presented system.

It is important to stress that the goal of this paper project is not to simply predict the speech recognition results; that is done more effectively using regression techniques

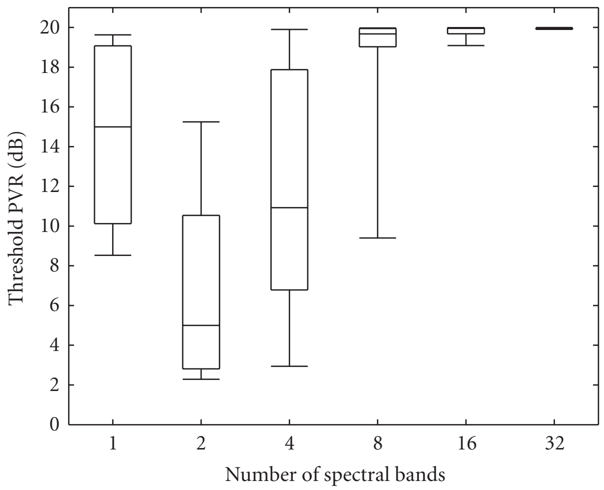


Figure 1: Results of the spectral discrimination test. Each box shows the median, quartiles, and 10th and 90th percentiles among the CI users for a specific stimulus type. The test and retest thresholds have been averaged.

such as support vector machines or neural networks. The goal is rather to explain the results in terms of equivalent signal processing, which then can be modified in order to simulate the effects of different signal processing techniques. The end goal is to be able to predict how an individual's speech recognition performance will be affected by a change in the speech processing algorithm, allowing for automatic optimization of speech recognition performance. Although such simulations may not be reliable enough for clinical use, they may give indications on the types of strategies that should be tested clinically. The general technique of fitting hearing instrument parameters by optimizing an objective measure of speech recognition ability has been proven useful in hearing aid fitting, for example, NAL-NL1 [13]. In fact, the approach dates back as far as the 1940s [14]. It is reasonable to believe that similar methods could be derived for cochlear implants in the future. Another goal is to show that a large fraction of the variance in speech recognition ability in CI users can be explained by spectral resolution properties.

2. Experimental Data

The presented framework was designed to simulate two psychophysical experiments: a spectral discrimination test and a sentence recognition test known as Hagerman's sentences [15]. These experiments are described below. 32 CI users performed the spectral discrimination task on two separate occasions (referred to as the test and retest) and the speech recognition task on one occasion. All listening tests were performed monaurally. Any hearing aid or CI on the other ear was turned off or removed. All participants were above 15 years old and had no known neurological disorders. 5 participants had severe prelingual hearing loss. The implants types were Med-El (17 users), Nucleus (12 users), and Clarion (3 users). The stimuli were presented in free field in a low-reverberant room at a fixed sound level of 70 dB SPL.

2.1. Spectral Discrimination Test. In this test, the listener's task is to determine which out of three consecutive stimuli differ from the other two (known as a 3 interval, 3 alternative forced choice task, or 3I3AFC). The signals are 200 milliseconds noise bursts which have been filtered so that the spectral density is matched to the long-term average of speech. The signals are then filtered into K bands between 200 and 8000 Hz that are evenly spaced on the ERBN scale [16]. For each interval, either the odd $(1, 3, 5, \ldots)$ or the even $(2, 4, 6, \ldots)$ bands are attenuated, and the attenuation is called the peak-to-valley ratio (PVR). The listener chooses the stimulus that was filtered differently from the other two stimuli. A modified up-down procedure (2-down, 1-up) was used to estimate the threshold PVR, which is defined as the 70.7% point on the psychometric function [17]. If the required number of correct responses is not found at the highest allowed PVR (20 dB), the fraction of correct responses at this level is recorded. The process is repeated for $K = 1, 2, 4, 8, 16$, and 32 bands. The single band case is simply an intensity discrimination task. Figure 1 shows the distribution of results in the test group as percentiles. The experiment is described in more detail in [5].

2.2. Hagerman's Sentences. The Hagerman's Sentences is a Swedish word recognition test which is used to determine speech recognition thresholds in noise. Similar tests exist in several languages. The test consists of 50 phonetically balanced words, organized into 5 sentence positions with ten possible words for each position, so that choosing one of the ten words for each position generates a grammatically correct but semantically meaningless sentence. After a sentence is presented along with speech-shaped masking noise, the listener's task is to repeat each sentence. The experimenter counts the number of correctly repeated words. The noise level is then shifted adaptively to converge to the level where 2 out of 5 words are correct, that is, 40% correct [18]. The corresponding SNR is noted as the result. These results can be viewed in Figure 7 along with model predictions.

3. Modelling Methods

To predict speech recognition performance, an entire speech communication chain is simulated: going from a speaker, via a medium, to a listener. The speaker tries to convey a sequence of words W through articulation, thereby generating the acoustic signal X. When the signal reaches the listener, it may be contaminated by additive noise and reverberation. In the CI case, the acoustic signal is transduced into electrical impulses, which give rise to a neural pattern sequence R. The listener's brain then acts as a classifier, finding an estimate $\widehat{W}$ of the word sequence. The speech recognition performance is defined as the fraction of words correctly identified. The following sections describe the process steps in more detail. Similar approaches are used in [11, 19] to estimate speech recognition performance for listeners with normal and impaired hearing.

3.1. Model of Speech Production and Recognition. Speech is inherently random; a single word will never be uttered exactly the same twice. Therefore, probabilistic models are appropriate to represent a speech source. Hidden Markov Models (HMMs) are used extensively in automatic speech recognition systems (and also to some degree in speech synthesis) due to their ability to model signals with variability in both duration and spectrum [20]. Here, each word in the speech recognition test is modelled by an HMM. The speech features used to train the models are calculated as follows: the time-domain signal is divided into 20 milliseconds, 50% overlapping frames. The spectrum of each frame is computed in 32 equally spaced bands on the ERBN scale from 300 to 8000 Hz, which is the frequency range encoded by the CIs used in the listening tests. The spectral magnitudes were then converted to dB. This representation is relevant, because Euclidean distance in this space is approximately proportional to perceptual difference for normal hearing listeners [21]. It is also quite similar to the signal encoding used in most CIs, which is useful in the individual adaptation procedure (see Section 3.4).

The speech features of the entire corpus are used to train a Gaussian Mixture Model (GMM) [22] with 40 components, which should correspond to approximately one component per phoneme of the Swedish language. Then, a 7-state HMM is trained on each single word, approximately one state per phoneme. The GMM is used as the output distribution for all word HMMs, and only the mixture weights are adapted for each state. This is known as a tied-mixture HMM [23], which reduces the degrees of freedom in the model dramatically compared to a standard continuous HMM. This approach reduces the risk of overfitting the model, which is useful here as only one utterance of each word is available in the Hagerman test. Finally, the single word HMMs can be concatenated to form a sentence HMM.

3.2. Word Classification. The inner workings of human speech perception is largely unknown, even after decades of research. The best automatic speech recognizers are still not close to human performance, which suggests that human speech recognition is, if not optimal, then at least not far from it. Therefore, the human brain is approximated by an optimal classifier. Here, "optimal" should be interpreted as classifying with minimal probability of error, which is achieved by choosing the most probable word sequence. After observing the neural pattern r, the listener finds the word sequence $\widehat{w}$ that maximizes the conditional probability,

$$\widehat{w} = \arg\max_{w} P_{W\mid R}(w \mid r). \tag{1}$$

In the case of Hagerman's sentences, every word sequence has equal prior probability. Therefore, the maximum a posteriori choice in (1) is identical to the maximum likelihood choice,

$$\widehat{w} = \arg\max_{w} f_{R\mid W}(r \mid w), \tag{2}$$

where $f_{R|W}$ expresses the probability density function (pdf) of the neural patterns elicited by a given word sequence. It is defined by the sentence HMM, and the search over all possible word sequences is efficiently computed using the Viterbi algorithm. The probability of correctly recognizing a word P_c can then be estimated as an average, by drawing samples w_i from the random variable W, giving

$$P_c = \frac{1}{N_w} E\left[\delta\left(W, \widehat{W}\right)\right] \approx \frac{1}{N_w M}\sum_{i=1}^{M}\delta(w_i, \widehat{w}_i). \tag{3}$$

Here, $E[\cdot]$ denotes the expectation over all possible sentences, N_w is the number of words in each sequence and $\delta(\cdot,\cdot)$ expresses the number of identical words in the two sequences. This procedure is quite similar to what would be done in a psychophysical experiment; it is simply averaging the number of correct answers. Unfortunately, a fairly large number of iterations M may be needed to find a good estimate, which can be quite time consuming if elaborate models are used. A more efficient approach is to consider the mutual information (MI) between the source W and the observation R, which expresses the amount of information available to the classifier:

$$I(W;R) = E_{R,W}\left[\log\left(\frac{f_{R|W}(R \mid W)}{f_R(R)}\right)\right]. \tag{4}$$

Like (3), this quantity must also be estimated using a Monte Carlo approach; however, it will converge faster, since each iteration yields a continuous estimate, compared to the binary result of 3. The MI is estimated using the following expression, where M is set to achieve the required accuracy

$$I(W;R) \approx \frac{1}{M}\sum_{i=1}^{M}\log\left(f_{R|W}(r_i \mid w_i)\right) - \log\left(f_R(r_i)\right), \tag{5}$$

where w_i is a word sequence drawn randomly from the distribution of W and r_i is the corresponding observed neural pattern.

The relation between MI and minimal classification error is given by rate-distortion theory, and is calculated using the Blahut algorithm [24]. In the case of Hagerman's sentences, 0.45 bits/word is needed to achieve the threshold 40% word recognition.

3.3. Models of CI Hearing. For the purposes of this paper, two models of CI hearing have been implemented. Model A is the simplest possible model that can account for the results of the spectral resolution test, while Model B is an attempt to model the actual signal transformations, including the CI speech processor, the electrical transduction from implant to auditory nerve, and the response of the auditory nerve cells. Model A has the advantage that its structure is very flexible and can mimic any result of the spectral discrimination task. Model B is less flexible, but since it is modelled after a physical cochlea, it can capture the effects of, for example, moving the electrodes. The following sections describe the models in detail.

3.3.1. Model A: Functional Signal Processing Model. Model A is an attempt to construct the simplest possible explanation

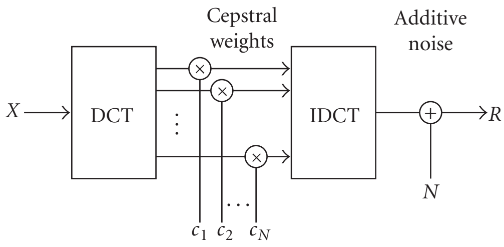


Figure 2: Block diagram of model A. The processing is applied to each time frame in the sequence.

of the observed psychophysical results. According to Occam's razor, one should always pick the simplest model when all other things are equal (a good discussion of this preference is provided in [25]). In the case of cochlear implants, we do have some knowledge of how a particular acoustic input gets encoded in the auditory nerve, but since that knowledge is limited, any model of that entire procedure is bound to be speculative. Therefore, model A is a good complement to the more realistic model B.

To create a simple functional model, it is first observed that the spectral shapes to be discriminated in the spectral discrimination test are periodic modulations on the ERBN scale. These are quite similar to the basis functions of mel-frequency cepstral coefficients (MFCCs), which is a ubiquitous signal representation in automatic speech recognition (this is of course not a coincidence, the experiment was designed with that in mind). As previously discussed, the incoming acoustic signal X is represented by a sequence of short-time auditory spectra in log units (i.e., dB). By applying the Discrete Cosine Transform (DCT) to each frame, the resulting features are effectively MFCCs. Each coefficient is then multiplied by a weight c_i, which controls the sensitivity to the corresponding spectral shape. Thereafter the feature vectors are inversely transformed, and Gaussian noise with unity variance is added to each coefficient. A block diagram of this process is shown in Figure 2.

This model is able to simulate any result of the spectral discrimination experiment, by adapting the weights $c = \{c_1 \cdots c_N\}$. For example, by attenuating the higher cepstral coefficients, the spectrum is smoothed, simulating a loss in frequency selectivity. The noise amplitude is adapted indirectly, by allowing all signal weights to scale.

3.3.2. Model B: Biologically Inspired Model. In this model, each user's implant settings were simulated, including frequency bands, "map-law", and (approximate) electrode positions. However, any additional signal processing being performed by the speech processor could not be modelled as these algorithms are proprietary. Instead, it has been assumed that the gain in each channel is mapped directly from the signal energy in the corresponding frequency band. Also, the signal preprocessing is implicitly modelled by assuming that the settings are ideally fitted, so that the signals utilize the full dynamic range in each channel (explained in more detail below). This mirrors the purpose of a front-end slow-acting compressor and a pre-emphasis filter. Thereafter, the spread of electric current in the cochlea is simulated in a 3D finite element model, implemented according to the specifications supplied by Rattay et al. [26]. The simulation was performed using the COMSOL Multiphysics finite element software [27].

A healthy cochlea contains approximately 32 000 auditory nerve fibers [28]. To model each nerve fiber individually is neither feasible nor necessary to capture the statistics of the neural activity. Instead, the auditory nerve is divided into tonotopic groups, and the expected spike rate from each group is computed. It is assumed that the perception of spectral shapes is based solely on the number of neural spikes from each group of neurons, that is, spike timing is neglected and the spike count is summed for each group and time frame. Furthermore, it is assumed that the summed response from each neural group is a monotonic function of the sum of the "activation" (defined below) in the nerve tissue elicited by each electrical impulse. These simplifications, although somewhat crude, are needed to make the speech recognition simulation possible.

The 3D model comprises one and a half turns of the cochlea. The electric field in the cochlea is examined in 30 degree steps, making for a total of 19 modelled fibers. Each modelled fiber is then assumed to be typical for the nerve fibers surrounding it, forming a "neural group" with homogenous properties. Each nerve fiber model consists of ten straight line sections, as illustrated in Figure 3. Unit monopolar impulses are simulated from electrodes at each of the corresponding 30 degree positions, and the activation as defined in [29] (the second spacial derivative of the electric potential) is recorded for each nerve segment. These results are stored in the activation matrices A_j, where $A_j(k,i)$ represents the activation in the kth section of nerve fiber i due to a unit impulse from electrode position j. From these data, responses to stimulation from electrodes at arbitrary positions could be found using linear interpolation.

In this model, it is assumed that all variation in frequency selectivity that is not explained by known factors (such as CI properties and settings) is due to degeneration of the auditory nerve. To model different stages of neural degeneration, the parameter c_k represents the fraction of nerve fibers that have the kth section, as well as all sections on the central side, intact. In other words, it is the fraction of fibers that are able to transmit an action potential when stimulated at section k. It is assumed that the degeneration is uniform across all neural groups. Studies on cochlear nerve degeneration show that this is generally not the case [30], but it may be an acceptable approximation, given that the spectral discrimination test used in this study does not give any frequency specific information. The weights c are used to compute the *transfer matrix* T, where each element T_{ij} represents the activity in neural group i due to a unit impulse from electrode position j, and is computed as

$$T_{ij} = \sum_{k=1}^{10} c_k A_j(k,i). \qquad (6)$$

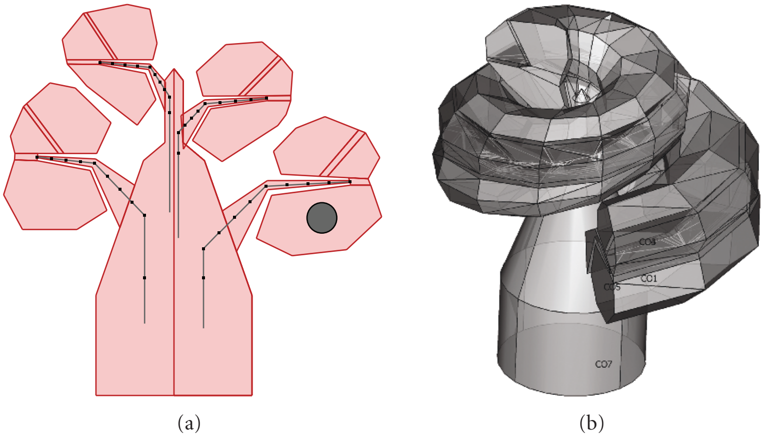


Figure 3: Visualization of 3D cochlear model. (a): Cross-section view of the model. Nerve fibers are shown as gray lines, with black dots at the boundaries between sections. One electrode positioned in the Scala Tympani is shown as a circle. (b): External view of the model cochlea.

Thus, the T matrix is used to model the spread of excitation in the cochlea. Examples of A and T matrices are shown in Figure 4.

The random variable R_t representing the neural feature vector at time t is computed as

$$R_t = g(TX_t' + N_t), \qquad (7)$$

where X_t' is the output from the CI electrode, N_t is isotropic Gaussian noise, and $g(\cdot)$ is a sigmoidal function applied to every element in the vector, defined as

$$g(y) = \frac{1}{1 + e^{(m-y)/s}}, \qquad (8)$$

where m is the midpoint of the sigmoid and s is its slope parameter. The sigmoid function is used to map the activation to neural firing rate, which has been shown to have a sigmoidal relation in single fiber measurements [31]. The values of the sigmoid parameters are not known; instead, it is assumed that the CI output range has been fitted to match these parameters, so that the neural "input" $TX_t' + N_t$ has a mean and standard deviation that is approximately m and s respectively, for each neural group. For simplicity, all noise sources (noise due to stochastic neural firing, decision noise, etc.) are modelled by the additive term N_t. A schematic view of the signal processing in the model is shown in Figure 5.

The model is adapted to individual data by varying the weights c and scaling the variance of the additive noise term N_t. With these free parameters, it is in general not possible to adapt the model to match a set of spectral discrimination results exactly. Therefore it is adapted using a maximum likelihood approach that is described in Section 3.4.1.

3.4. Adaptation to Psychophysical Data. The models of CI hearing have a number of individual (free) parameters, which vary between subjects. Some parameters can be observed directly (such as the number of active channels), while others cannot (such as the degeneration of the auditory nerve). For each subject, the latter set of parameters are estimated by matching the results from the spectral discrimination experiment described in Section 2.1 in a simulation using a Maximum Likelihood criterion, as described below. The estimated model parameters are then applied in the speech recognition simulation to predict the speech recognition threshold. The full procedure is illustrated in Figure 6.

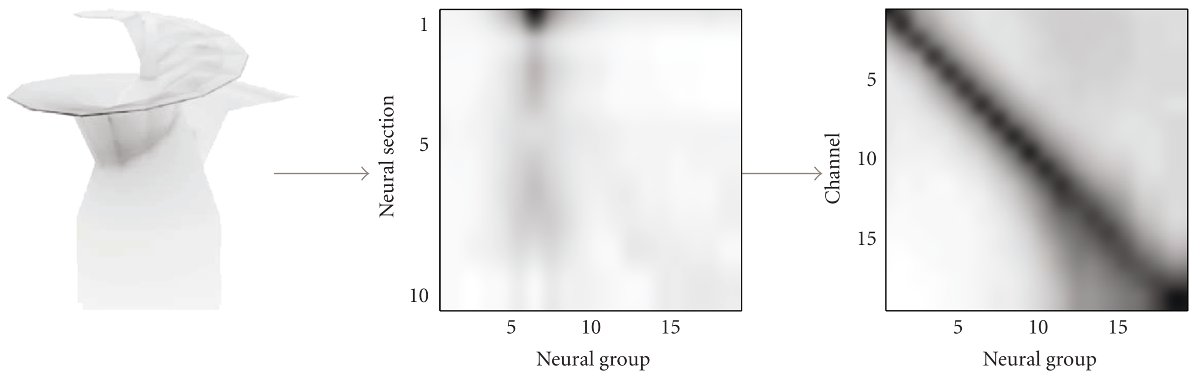
3.4.1. Maximum Likelihood Adaptation. The models of hearing have limited degrees of freedom, and it is generally not possible to match the experimental thresholds exactly. However, the measured thresholds should not be interpreted as deterministic; all psychophysical results have an inherent uncertainty due to the probabilistic nature of perception. In the spectral discrimination test, the experimenters calculated a threshold estimate $\hat{m}_i$ and associated error variance σ_i^2 for each condition. From these data, a probability distribution of the true thresholds m_i can be formed (assuming estimation errors are independent and Gaussian):

$$f(m_1 \cdots m_N) = \prod_i f(m_i) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(m_i-\hat{m}_i)^2/2\sigma_i^2}, \qquad (9)$$

where $f(\cdot)$ denotes a probability density (the likelihood function). To find the optimal parameter set, an iterative optimization scheme is employed (the Nelder-Mead simplex method, a popular "hillclimbing" algorithm [32]). The algorithm converges to the set of parameters that (locally) maximizes the likelihood of the simulated thresholds.

4. Results

In order to predict the speech recognition threshold in noise (SRT) of a specific user, synthesized speech signals were mixed with noise and run through the user-adapted CI model. The signal-to-noise ratio (SNR) was initiated at 20 dB and the information rate was estimated according to 5. The estimate was refined until the threshold rate 0.45 bits/word was outside the 99% confidence interval. The SNR was shifted in 2 dB steps until two consecutive SNR levels were found to be on either side of the threshold rate. The SRT



(a) (b) (c)

Figure 4: Results from the electric field simulation. The activation over the cochlear nerve tissue during a unit impulse from electrode position 7 (a) is mapped to the activation matrix A_7 (b). A matrices are combined to form the matrix T (c), here with $c_k = 1$ for all k. The electrodes and neural groups are indexed in basal to apical direction and the neural sections in peripheral to central direction.

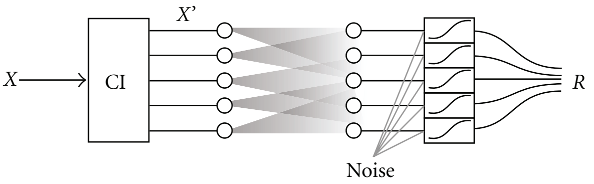


Figure 5: Block diagram of Model B. The CI unit provides a filter bank and sound level to electric current mapping. The current from each channel is spread to the neural groups according to the transfer matrix T. Additive gaussian noise is used to simulate the variability of the neural response.

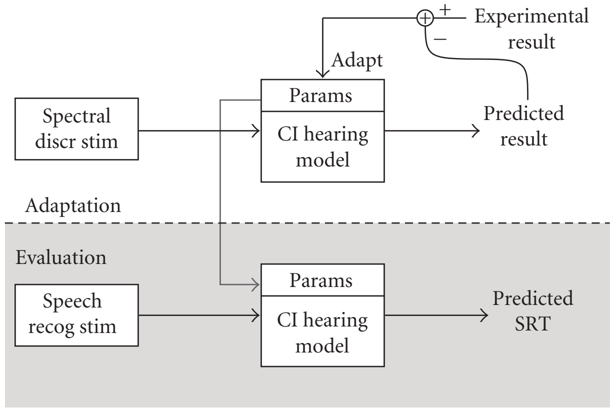


Figure 6: Adaptation and prediction procedure. The parameters of the CI model are adapted to match a set of experimental results, in this case the spectral discrimination thresholds. The adapted parameters are then used in a simulated speech recognition test.

Table 1: SRT prediction accuracy. The median absolute errors and Spearman correlation coefficients with corresponding P-values for SRT predictions using all combinations of models and data sets, including the average of the test and retest predictions (denoted "avg"). The table also includes prediction results from a jackknife linear regression analysis (noted "R").

Model/data	Error (dB)	Correlation	P-value
A/test	5.0	0.57	$6 \cdot 10^{-4}$
A/retest	4.0	0.73	$3 \cdot 10^{-6}$
A/avg	4.4	0.72	$4 \cdot 10^{-6}$
B/test	4.3	0.52	$2 \cdot 10^{-3}$
B/retest	4.4	0.72	$4 \cdot 10^{-6}$
B/avg	4.8	0.68	$2 \cdot 10^{-5}$
R/test	3.3	0.54	$2 \cdot 10^{-3}$
R/retest	4.0	0.62	$2 \cdot 10^{-4}$
R/avg	2.6	0.66	$4 \cdot 10^{-5}$

estimate was then found by linear interpolation between the two levels. The process was repeated for test and retest data, for both CI models, and for each individual CI user. The results are displayed in Figure 7. Correlations between predictions and experimental data are tabulated in Table 1, along with median absolute prediction errors. As a reference, prediction using jackknife (leave one out cross-validation) linear least squares estimation (LLSE) or commonly known as linear regression has been performed, noted in the table as model "R". This method finds the linear dependency between the spectral discrimination and speech recognition results that minimizes the mean square prediction error. Statistical analysis using Spearman's rank correlation test has showed that all the measured correlations are significant at the 99% confidence level. It can be noted that the proportion of the variance in the data that is "explained" by the prediction is equal to the square of the correlation coefficient, which means that models A and B explain more than half of the variance when fitted to the retest data.

5. Discussion

The results in Figure 7 and Table 1 show that the two CI models give predictions that are approximately equally accurate, but not identical. Both models also overestimate performance (i.e., underestimate the SRT), which most likely

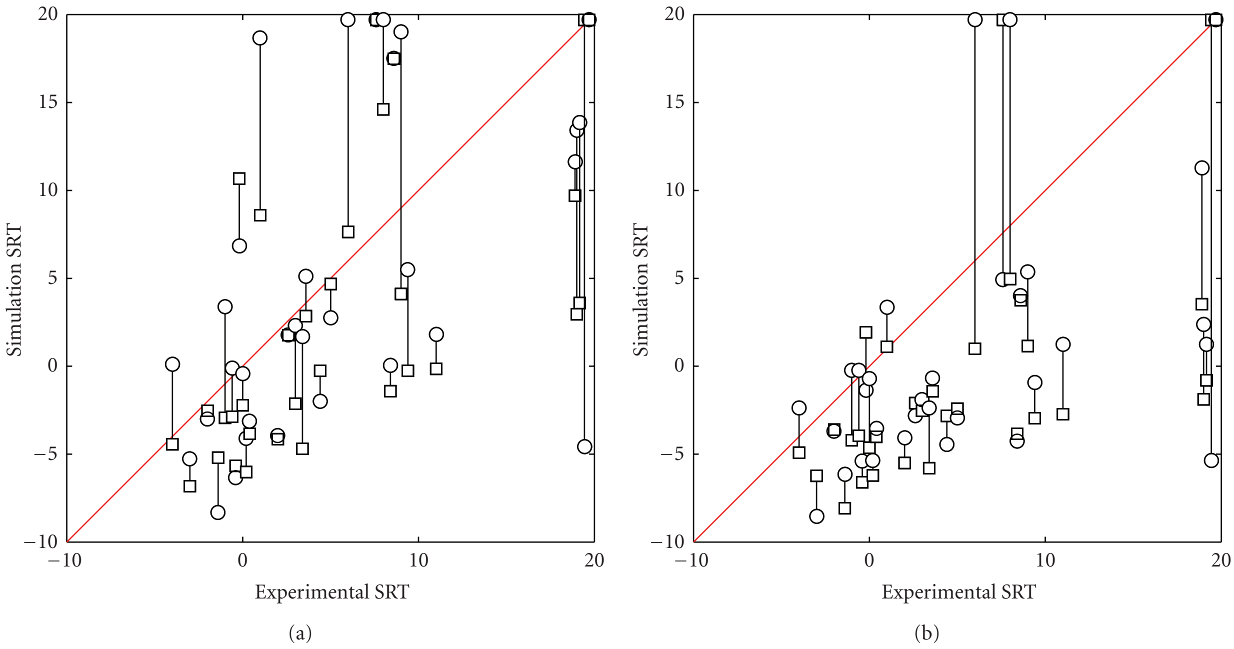


Figure 7: Speech recognition threshold for 32 CI users, as measured experimentally and as predicted by models A (a) and B (b). Each line connects the SRT predicted from the test (circles) and retest (squares) data. The SRT is expressed in dB and a lower value means better speech recognition.

is a result of using an optimal decision unit as a model of the cognitive process. For many subjects there is a large discrepancy between the test and retest predictions, which indicates that the spectral discrimination test results are not very reliable. The tabulated correlation coefficients suggest that the retest measurements are more accurate, which is reasonable; some users may be confused by the fairly abstract discrimination task on the first occasion. The two models give different predictions, but are approximately equally accurate. The main advantage of model A is its ability to simulate any spectral resolution exactly, which model B cannot. However, model B is more useful because it is able to simulate different types of CI signal processing and coding strategies, as discussed below. The regression predictor gives smaller errors on average, as there is no bias present; looking at the correlation coefficients, however, the regression predictor is comparable to the models A and B.

One major weakness of the spectral discrimination experiment is its inability to assess frequency specific information. This is because the signals used in the experiment are modulated across the entire frequency range. Thus, any frequency specific deficiency, for example, hearing only a small range of frequencies, cannot be detected. However, this does not imply that the resulting prediction will be inaccurate; as mentioned earlier, the stimuli are similar to MFCCs, which in some sense are the “basis functions” of speech, and one may argue that it is a more relevant domain than the frequency domain when modelling speech recognition. Of course, the optimal case would be to have access to data in both domains. Another possible weakness in the data is that the experiments were not done using roving sound levels, which means that discrimination can potentially be performed by monitoring the intensity in a narrow frequency range, a problem that was illuminated in a recent study [33]. In this particular case, it should be a minor concern; since the data include a 1-band case (i.e., intensity discrimination), the models can adapt to situations where listeners are using mainly intensity as a cue for discrimination and this will be reflected in the speech recognition simulation.

Although many simplifications have been made even in our advanced model B, it is fairly unlikely that using more accurate modelling would improve the prediction to any significant degree; there are simply too many unmeasurable factors influencing the outcome. However, the cause of an individual discrepancy may be unveiled in further psychophysical experiments. An interesting next step would be to construct a *ladder of ecological validity*: a series of experiments ranging from simple spectral discrimination up to sentence recognition, having intermediate steps that are more controlled than speech but more realistic than filtered noise. In this way, one is able to check which of the underlying assumptions in the presented framework hold and which do not.

A quite useful application of this framework would be to modify the signal processing and coding strategy of a modelled implant and observe the impact on the predicted speech recognition threshold. In this way, novel schemes can be evaluated by simulation. Although such results will never be a replacement for human trials, it can be useful at the development stage, as human trials tend to be very lengthy. Since the CI models are individualized, it would be possible to estimate which strategy is most suited for that particular user.

6. Conclusion

The two presented models of CI hearing are both capable of explaining as much as 50% of the variance in speech recognition capability among CI users. The framework for simulating speech communication is useful for evaluating novel signal processing strategies for CIs, both for the CI users in general and for finding optimal settings for individual users. The spectral discrimination test used to assess users' hearing capabilities has shown to be useful, even though additional listening tests might enable more accurate modelling and prediction.

References

[1] F.-G. Zeng, A. N. Popper, and R. R. Fay, Eds., *Cochlear Implants: Auditory Prostheses and Electric Hearing*, Springer, New York, NY, USA, 2004.

[2] R. V. Shannon, "Multichannel electrical stimulation of the auditory nerve in man—I: basic psychophysics," *Hearing Research*, vol. 11, no. 2, pp. 157–189, 1983.

[3] L. M. Friesen, R. V. Shannon, D. Baskent, and X. Wang, "Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants," *Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1150–1163, 2001.

[4] B. A. Henry and C. W. Turner, "The resolution of complex spectral patterns by cochlear implant and normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 113, no. 5, pp. 2861–2873, 2003.

[5] E. Molin, A. Leijon, and H. Wallsten, "Spectro-temporal discrimination in cochlear implant users," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 3, pp. 25–28, Philadelphia, Pa, USA, 2005.

[6] Q.-J. Fu, R. V. Shannon, and X. Wang, "Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing," *Journal of the Acoustical Society of America*, vol. 104, no. 6, pp. 3586–3596, 1998.

[7] B. Lyxell, B. Sahlen, M. Wass, et al., "Cognitive development in children with cochlear implants: relations to reading and communication," *International Journal of Audiology*, vol. 47, supplement 2, pp. S47–S52, 2008.

[8] C. V. Pavlovic, G. A. Studebaker, and R. L. Sherbecoe, "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," *Journal of the Acoustical Society of America*, vol. 80, no. 1, pp. 50–57, 1986.

[9] T. Y. C. Ching, H. Dillon, and D. Byrne, "Speech recognition of hearing-impaired listeners: predictions from audibility and the limited role of high-frequency amplification," *Journal of the Acoustical Society of America*, vol. 103, no. 2, pp. 1128–1140, 1998.

[10] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

[11] S. Stadler, A. Leijon, and B. Hagerman, "An information theoretic approach to estimate speech intelligibility for normal and impaired hearing," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, Antwerpen, Belgium, 2007.

[12] M. S. A. Zilany and I. C. Bruce, "Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery," in *Proceedings of the 3rd International IEEE EMBS Conference on Neural Engineering*, pp. 481–485, Kohala Coast, Hawaii, USA, 2007.

[13] D. Byrne, H. Dillon, T. Ching, R. Katsch, and G. Keidser, "NAL-NL1 procedure for fitting nonlinear hearing aids: characteristics and comparisons with other procedures," *Journal of the American Academy of Audiology*, vol. 12, no. 1, pp. 37–51, 2001.

[14] W. Radley, W. Bragg, R. Dadson, et al., *Hearing Aids and Audiometers*, Medical Research Council Special Report Series no. 261, Her Majesty's Stationery Office, London, UK, 1947.

[15] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scandinavian Audiology*, vol. 11, no. 2, pp. 79–87, 1982.

[16] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[17] H. Levitt, "Transformed up- down methods in psychoacoustics," *Journal of the Acoustical Society of America*, vol. 49, no. 2, part 2, pp. 467–477, 1971.

[18] B. Hagerman and C. Kinnefors, "Efficient adaptive methods for measuring speech reception threshold in quiet and in noise," *Scandinavian Audiology*, vol. 24, no. 1, pp. 71–77, 1995.

[19] A. Leijon, "Estimation of secondary information transmission using a hidden Markov model of speech stimuli," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 423–432, 2002.

[20] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

[21] R. Plomp, *Aspects of Tone Sensation*, Academic Press, London, UK, 1976.

[22] G. J. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, New York, NY, USA, 2000.

[23] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2033–2045, 1990.

[24] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 1991.

[25] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, UK, 2006.

[26] F. Rattay, R. N. Leao, and H. Felix, "A model of the electrically excited human cochlear neuron—II: influence of the three-dimensional cochlear structure on neural excitability," *Hearing Research*, vol. 153, no. 1-2, pp. 64–79, 2001.

[27] Comsol Multiphysics, http://www.comsol.com.

[28] H. Spoendlin and A. Schrott, "Analysis of the human auditory nerve," *Hearing Research*, vol. 43, no. 1, pp. 25–38, 1989.

[29] F. Rattay, "Analysis of models for external stimulation of axons," *IEEE Transactions on Biomedical Engineering*, vol. 33, no. 10, pp. 974–977, 1986.

[30] C. E. Zimmermann, B. J. Burgess, and J. B. Nadol Jr., "Patterns of degeneration in the human cochlear nerve," *Hearing Research*, vol. 90, no. 1-2, pp. 192–201, 1995.

[31] C. A. Miller, P. J. Abbas, B. K. Robinson, J. T. Rubinstein, and A. J. Matsuoka, "Electrically evoked single-fiber action potentials from cat: responses to monopolar, monophasic stimulation," *Hearing Research*, vol. 130, no. 1-2, pp. 197–218, 1999.

[32] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder-Mead simplex method in low dimensions," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[33] M. J. Goupell, B. Laback, P. Majdak, and W.-D. Baumgartner, "Current-level discrimination and spectral profile analysis in multi-channel electrical stimulation," *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3142–3157, 2008.