Probabilistic modelling of hearing:
Speech recognition and optimal audiometry

Svante Stadler

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)

Stockholm 2009
Abstract

Hearing loss afflicts as many as 10% of our population. Fortunately, technologies designed to alleviate the effects of hearing loss are improving rapidly, including cochlear implants and the increasing computing power of digital hearing aids. This thesis focuses on theoretically sound methods for improving hearing aid technology. The main contributions are documented in three research articles, which treat two separate topics: modelling of human speech recognition (Papers A and B) and optimization of diagnostic methods for hearing loss (Paper C).

Papers A and B present a hidden Markov model-based framework for simulating speech recognition in noisy conditions using auditory models and signal detection theory. In Paper A, a model of normal and impaired hearing is employed, in which a subject’s pure-tone hearing thresholds are used to adapt the model to the individual. In Paper B, the framework is modified to simulate hearing with a cochlear implant (CI). Two models of hearing with CI are presented: a simple, functional model and a biologically inspired model. The models are adapted to the individual CI user by simulating a spectral discrimination test. The framework can estimate speech recognition ability for a given hearing impairment or cochlear implant user. This estimate could potentially be used to optimize hearing aid settings.

Paper C presents a novel method for sequentially choosing the sound level and frequency for pure-tone audiometry. A Gaussian mixture model (GMM) is used to represent the probability distribution of hearing thresholds at 8 frequencies. The GMM is fitted to over 100,000 hearing thresholds from a clinical database. After each response, the GMM is updated using Bayesian inference. The sound level and frequency are chosen so as to maximize a predefined objective function, such as the entropy of the probability distribution. It is found through simulation that an average of 48 tone presentations are needed to achieve the same accuracy as the standard method, which requires an average of 135 presentations.

**Keywords:** auditory models, probabilistic modelling, speech modelling, human speech recognition, hearing aids, cochlear implants, psychoacoustics, diagnostic methods, optimal experiments, audiometry.
List of Papers

The thesis is based on the following papers:


In addition to papers A-C, the following papers have also been produced in part by the author of the thesis:


Acknowledgements

The work behind this thesis has been carried out at the Sound and Image Processing Lab (SIP) at the Royal Institute of Technology (KTH) in Stockholm. It has been supported by research grants from the EU Hearcom project, the Swedish Research Council and GN Resound.

I consider myself lucky to have Arne Leijon as my supervisor. His knowledge and experience has led to me to wonderful insights about the nature of signals and probabilities. I have enjoyed all of our discussions, which always are intellectually stirring and rarely stays on topic. Your generous and supportive ways has made work feel a little bit less like work. I must also thank my co-supervisor Bastiaan Kleijn for his no-nonsense approach to research and problem solving, which I hope has rubbed off on me. To each and every person who works at SIP or has only passed by, thank you for providing a stimulating work environment.

I am grateful to Bert de Vries and Tjeerd Dijkstra at GN Resound in Eindhoven for their support and valuable feedback on our optimal audiometry project. Although it was only during a short time, it has been very productive and I have learned a lot from these few months.

I am indebted to Elisabet Molin and Björn Hagerman for kindly letting me access the data from their clinical experiments, and to Martin Eklöf for helping me get information about cochlear implant users. I must also thank Martin for many interesting discussions and I hope that we can collaborate on a new project very soon!

Finally, I would like to thank Ylva, the love of my life, for her continuous support and encouragement throughout every step leading to this thesis. I look forward to every new day with you and our wonderful daughter Blenda.

Svante Stadler
Stockholm, April 2009
Contents

I Introduction 1
  1 A probabilistic view of hearing ................. 2
    1.1 The psychometric function .................. 2
    1.2 Signal detection models ..................... 5
    1.3 Estimation of psychophysical thresholds .... 9
  2 Models of normal, impaired, and aided hearing .... 12
    2.1 Models of normal hearing ................... 12
    2.2 Models of impaired hearing ................. 15
    2.3 Cochlear implant models .................... 16
  3 Auditory information processing of speech .......... 17
    3.1 A model of speech communication ............. 19
    3.2 Probabilistic speech models ................. 19
    3.3 Estimation of speech intelligibility ......... 22
  4 Summary of Contributions ........................ 25
  5 Conclusions .................................... 27
References ........................................ 28

II Included papers 37
A An Information Theoretic Approach to Predict Speech Intelligibility for Listeners with Normal and Impaired Hearing A1
  1 Introduction ................................... A1
  2 Model of Speech Communication .................. A2
    2.1 Signal Representation ........................ A3
    2.2 Probabilistic Speech Model .................. A3
    2.3 Model of Noisy Speech ....................... A4
    2.4 Auditory Model .............................. A4
    2.5 Estimation of Information Transmission .... A6
  3 Comparison with Experimental Data .............. A6
  4 Discussion and Future Work .................... A8
  5 Conclusion .................................... A9
B Prediction of speech recognition in cochlear implant users by adapting auditory models to psychophysical data

1 Introduction .............................................. B1
2 Experimental data ....................................... B2
2.1 Spectral discrimination test ......................... B3
2.2 Hagerman’s sentences ................................. B3
3 Modelling Methods ..................................... B3
3.1 Model of speech production and recognition .... B4
3.2 Word classification .................................... B5
3.3 Models of CI hearing ................................. B6
3.4 Adaptation to psychophysical data ................. B10
4 Results ..................................................... B12
5 Discussion ................................................ B12
6 Conclusion ................................................ B14

C Bayesian Optimal Pure Tone Audiometry with Prior Knowledge

1 Introduction .............................................. C1
2 Method ..................................................... C2
2.1 Parameterization of the Psychometric Function ... C2
2.2 Prior Knowledge ....................................... C2
2.3 Bayesian Procedure .................................. C6
2.4 MAP Estimation ....................................... C11
2.5 Ordered Frequency Presentation ................. C11
2.6 Reference Method .................................. C12
3 Results ..................................................... C12
3.1 Objective Function Evaluation ..................... C13
3.2 Gaussian Approximation Evaluation ............. C13
3.3 MAP Estimation Evaluation ....................... C14
3.4 Evaluation of Ordered Frequency Presentation .. C15
4 Discussion ................................................ C15
4.1 Future Work .......................................... C17
5 Conclusion ................................................ C17
# Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>BM</td>
<td>Basilar Membrane</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>ERB</td>
<td>Equivalent Rectangular Bandwidth</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>IHC</td>
<td>Inner Hair Cell(s)</td>
</tr>
<tr>
<td>IIR</td>
<td>Infinite Impulse Response</td>
</tr>
<tr>
<td>JND</td>
<td>Just Noticeable Difference</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum A-Posteriori</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>OHC</td>
<td>Outer Hair Cell(s)</td>
</tr>
<tr>
<td>PF</td>
<td>Psychometric Function</td>
</tr>
<tr>
<td>SII</td>
<td>Speech Intelligibility Index</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SRT</td>
<td>Speech Reception Threshold</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Time Fourier Transform</td>
</tr>
</tbody>
</table>


Part I

Introduction
Introduction

Our senses are our interface to the world; we need to see, hear, smell and touch our surroundings constantly to function correctly. While vision is our main way of navigating this world, hearing is essential to speech communication, which we use to form and maintain social bonds with the people around us. Hearing loss, complete or partial, can lead to considerable social stigma and alienation, which may influence the overall mental health and even lead to depression [5]. A Swedish survey have shown that about 10 percent of the population report having some form of hearing loss [74]. In many cases, hearing aids provide a great relief, allowing for not only improved hearing, but an overall increase in quality of life [13]. In the last decade, improvements in cochlear implant (CI) technology have enabled people who were previously deaf or near deaf to engage in conversation, even in moderately noisy environments [53]. For those of us blessed with normal (unimpaired) hearing, we are equipped with an extremely sensitive and robust instrument. Our ears have a dynamic range that is unrivaled by any other sensory modality, with a factor of $10^{12}$ between the faintest and loudest sound intensities we can perceive. Perhaps most importantly, we are able to recognize speech in extremely noisy environments - an ability that, at the time of writing, surpasses that of automatic speech recognition (ASR) systems by a fair margin, even after 50 years of efforts to construct such systems [89].

One of the main goals of speech perception research is to be able to quantify the intelligibility of noisy speech objectively, without the need for human listening tests. There are many available methods, some of them with very impressive results. These methods are generally based on ad hoc reasoning, relying on ’rules of thumb’ rather than a mathematically rigid theoretical foundation. This can of course be justified by referring to how little is actually known about human perception; we simply have too little knowledge to base our theories on, so we might as well use any model that fits our data reasonably well. But if our goal is to one day fully understand the hearing system, we should do so by relying on theoretically sound principles. Probability theory and information theory provides that framework on which a quantitative model of human communication can be
The use of information theory to make inference about human perception is a highly functional approach; the assumed model structures may not be biologically plausible. But at a high level of abstraction, there is a form of correspondence, as humans evidently are good at transmitting information using speech. In fact, this is done in such a robust manner that it may be a reasonable approximation that our brains make optimal decisions, utilizing all available information and thereby obtaining the theoretical performance bounds imposed by information theory. Indeed, a growing number of researchers, particularly in the visual perception field, are recognizing that humans are approximately Bayesian estimators, integrating sensory data with prior knowledge in an optimal way [26,83].

The abstractions of information theory also naturally lend themselves to describe diagnostic methods; in psychophysical experiments, the goal is to extract as much information about the subject as possible using as little time and resources as possible. In this thesis, a novel method for estimating sensory thresholds in a series of experimental trials is presented, where the expected information gain in each trial is maximized. Although the method can be used to find any sensory threshold, it is demonstrated on the application of finding the threshold for pure tones presented in quiet.

1 A probabilistic view of hearing

Sensory experiences can be viewed as a way of making measurements about the world around us. However, the measurements are not perfect and there are of course limits to what we are able to hear. An integral part of psychoacoustical research is to quantify these limits. This line of research was initiated by Gustav Theodor Fechner in his *Elemente der Psychophysik*, published in 1860. A central concept in Fechner’s work was the *sensory threshold*, the pivotal point along a physical dimension that separates detectable from undetectable stimuli. The notion of a sensory threshold supports the idea that there is a discontinuity in perception, that sounds above a threshold are always audible and the sounds below it are completely inaudible. Although this may be true on a macro scale (i.e. far above and far below the threshold), experiments in all aspects and modes of perception show that this is not the case for stimuli that are near the threshold.

1.1 The psychometric function

Fechner himself noted that when stimuli were presented near the threshold repeatedly, subjects tended to give positive responses only a fraction of the time, and that this fraction increased towards 1 as the stimulus level increased. For this reason he introduced the *psychometric function* (PF),
which expressed the probability of a positive response as a function of the stimulus level. However, Fechner attributed this “softness” of the PF to time-varying sensitivities in the subject (from trial to trial), and this variability manifested itself as the slope of the PF. Therefore, the belief was still held that the threshold was a sharp transition, that stimuli below it had no impact on the listener. Since then, evidence has mounted showing that there is no discrete threshold; the threshold can only be interpreted as a stimulus level corresponding to some predefined probability of detection. This also resonates with findings from neuroscience showing that nerve cell activity is essentially random. Thus, human perception should be viewed as probabilistic. This point is argued empathically in [30] and expanded upon in Section 1.2.

**Parametric representation**

The PF, when expressed as a function of a logarithmic unit such as dB, can generally be assumed to have a “sigmoidal” shape that is monotonic (an increase in level does not decrease detection probability) and have a lower and upper asymptote. As such, it can be expressed as

\[
\Psi(x) = \gamma + (1 - \lambda - \gamma)p\left(\frac{x - m}{s}\right),
\]

where \(\gamma\) is the lower asymptote (the false positive probability, or guess probability in a forced choice experiment), \(\lambda\) is the probability of “lapsing” (missing a stimulus that is far above the threshold), and \(p(\cdot)\) is a sigmoidal function such that \(p(-\infty) = 0, p(0) = 1/2,\) and \(p(\infty) = 1.\) There are various functions that are used for this purpose, including the cumulate of any bell-shaped probability distribution such as Gaussian. The choice of function makes for very subtle differences in the PF; the choice is commonly made out of mathematical convenience. In Paper C, the choice fell upon the cumulative Gaussian distribution, for compatibility with the normal distribution used for the threshold parameter.

The parameters \(m\) and \(s\) are used to control the threshold position and slope respectively. Here, \(m\) is called the primary parameter, because it is the parameter that is of primary interest in clinical measurements such as pure tone audiometry. It indicates the steepest portion of the PF, and is thus a reasonable definition of the threshold.

\(\gamma, \lambda, \) and \(s\) are here referred to as the secondary parameters, as they are generally of little clinical interest. However, they serve an indirect purpose: the PF can be used to estimate its own parameters in a Bayesian update framework, as described in Paper C. The values of the secondary parameters will limit the amount of information that can be gained from each trial, as shown in Figure 1. As the figure shows, the information gain decreases when the parameters are increased. The optimal 1 bit per trial can only
be achieved when all secondary parameters are zero. In the case of \( N \)-alternative forced choice, \( \gamma = 1/N \) (the probability of guessing correctly). This means that a in a 2AFC experiment, a maximum of 0.54 bits can be gained from each trial, which explains why the method gives poor estimates compared to a classical yes/no method when the number of trials is small [56].

![Graph](image)

**Figure 1:** Maximum expected information gain per trial in a psychophysical detection experiment, as a function of PF parameters \( s \) (slope) and \( \gamma \) (false positive probability). The gain is calculated for a Gaussian prior distribution with unity variance. Due to the symmetry of the PF, the lapse probability \( \lambda \) has the same impact on the information gain as \( \gamma \). The maximum of 1 bit per trial can only be achieved when all secondary parameters are zero. The primary parameter \( m \) does not influence the information gain.

Many sensory thresholds that are determined in experiments are of mainly academic interest; it is not clear whether an individual’s ability to discriminate pure-tone frequencies (as an example) will affect her daily life in any way. However, this thesis focuses on what is perhaps the two most important psychoacoustic measurements: the Speech Reception Threshold (SRT) in noise and the pure-tone hearing thresholds (also known as the audiogram).
1.2 Signal detection models

The assumption in the signal detection models of human perception is that at the end of the perception chain is a decision device, which may or may not be ideal/optimal. The device is either a simple detector (i.e. signal or no signal) or a classifier that chooses from $N$ possible responses, depending on the task at hand. The decision is made based on the information available to the detector. In the case of hearing, that information is not the acoustic stimulus, but rather the encoding of the stimulus by the auditory system. Such a neural encoding is sometimes known as a *sensory pattern*. Given the space of possible sensory patterns, the classifier can be completely specified using decision boundaries, that divides the sensory space into decision regions. Given a particular pattern, the device simply checks which region it is in. This is illustrated by Figure 2, which shows the decision regions of a vowel classifier using only the first two formants of the spectrum.

![Figure 2: An example of decision regions: vowel regions, as a function of the first two formant frequencies. The phonetic symbols show the centers of the regions, the archetypes for the vowels. The regions are constructed using a minimum squared error criterion. A human observer may use slightly different regions.](image)

The optimal detector

The definition of “optimal” may vary depending on the circumstances. For instance, there may be costs associated with certain decision errors, so that
avoiding those may be worth an increase in the total error rate. As an example, a doctor may choose to give medicine to a patient although the chance of the patient having the disease is below 50%, because the costs of giving medicine to a healthy patient is much smaller than not treating a patient who has the disease. This tradeoff is formalized by statistical decision theory, introduced by Wald in the 1940’s [92]. If all errors have the same cost, which is a reasonable assumption for speech recognition, the optimal choice is simply the most likely choice, known as the maximum a posteriori (MAP) choice\(^1\). The decision rule is derived in the following. Here, capital letters denote random variables, lower case letters represent observations, and \( P(\cdot) \) is a probability distribution (discrete or continuous).

### MAP classification

After observing a signal \( x \), the MAP classifier’s task is to find which out of a set of \( N \) signal sources was most likely to have generated the signal. For this to be possible, the classifier must have access to the probability distribution \( P_{X|D}(x|d) \), i.e. the probability that the source \( d \) would generate the signal \( x \), where \( d = 1 \ldots N \). This is commonly referred to as the likelihood function. If it expresses the true probability distribution that generated the data, then the classifier is optimal. However, in many real scenarios such as speech recognition, the true distribution is unknown and approximate distributions must be used. Once an appropriate likelihood function has been chosen, the posterior probabilities of the signal sources is easily calculated using Bayes’ rule,

\[
P_{D|X}(d|x) = \frac{P_{X|D}(x|d)P_D(d)}{P_X(x)}. \tag{2}
\]

The optimal decision is then found by choosing the signal source that maximizes the posterior probability,

\[
d_{\text{opt}} = \arg \max_d P_{D|X}(d|x) = \arg \max_d P_{X|D}(x|d)P_D(d). \tag{3}
\]

### The human detector

There is no question that the mathematical foundation of signal detection theory is sound; what is less clear is how well and under which circumstances the theory can be used to model human perception. This topic has been actively researched for many decades, resulting in large amounts of evidence supporting both sides of the argument. Here, the question is divided into three sub-questions, which are addressed in the following:

\(^1\)A similar and perhaps more common approach is the Maximum Likelihood (ML) method, which is a special case of MAP where all decision are assumed to have the same prior probability.
1. *Is the physiology of the auditory system compatible with the signal detection framework?* A basic premise of the signal detection approach is that the signal transduction stage and the decision stage involved in perception are completely separate. Although these stages are represented in different areas of the brain, it is not clear how separate the processes are; the presence of efferent nerves in the cochlea suggest that central processes are able to influence the cochlear activity [33], and there is some evidence showing that attention influences the auditory process [90]. The theory does not specify where to draw the line between feature extraction and classification, but looking at the structure of the auditory pathway (see Figure 3), it seems unreasonable to draw such a line at any point. A far more viable explanation is that each stage is processing the signal until only the relevant information (i.e. the decision) remains. However, the properties of the system may still be similar to that of a signal detection system, although the structure is very different.

2. *Can the model explain the findings of psychophysical experiments?* A consequence of using an optimal or near optimal detector in perception would be that cues from several independent perceptual “channels” could be involved in the detection of a stimulus. This is supported by vast experimental evidence showing that detection of “sub-threshold” stimuli is in general possible when multiple such stimuli are presented simultaneously, and the detection is in most cases equal or near that of an ideal detector. This has been shown for multi-tone stimuli [10, 29], binaural tones and noise [68], and to some degree even audiovisual cues [22].

3. *How optimal is the human detector?* The answer to this question is obviously dependent on the listening task; simple detection or discrimination tasks require very little cognitive processing and it seems that humans can operate close to optimality on such tasks, in the sense that we are able to adjust our response criterion to the prior probability of a stimulus in a near optimal way [85]. On the other hand, our cognitive ability is limited by a number of biases and shortcomings such as limited memory span and phonological skills. The perhaps most challenging auditory task is that of speech recognition. It is not known how close humans are to optimal detection, as the optimal detector is not well defined. Studies must therefore focus on the differences between listeners. It has been found that recognition of full sentences is more cognitively demanding than single word recognition [28]. Several cognitive factors have been identified that correlate with speech recognition ability [72]. Also, there are indications that hearing impaired persons may have some type of cognitive deficit that reduces their ability to recognize noisy speech [43].
Figure 3: A simplified view of the organization of the auditory pathway in the brainstem, adapted from [65]. The auditory cortex can be divided into structural components in a similar way. It is not clear which components are to be considered signal processors and which are part of the decision device.

**Noise**

Central to the signal detection models is the presence of noise or inaccuracies; if there are no noise sources, the system should have infinite accuracy, which is clearly not the case with neither human nor non-human (existing) detectors. One obvious source of noise in hearing is the excitation of the auditory nerve fibers, which is stochastic in its nature – either there is an action potential, or there is not, and the probability of firing is a function of the stimulus intensity. However, in simulations of various psychoacoustic experiments where an ideal detector operates on the (simulated) auditory nerve response, it has been found that such systems greatly outperform normal hearing humans [40, 79]. Therefore, one can conclude that there are noise sources in the higher levels of perception, and since the entire auditory system is composed of nerve cells, it seems reasonable to assume
that similar stochastic properties innervate all levels of the system – noise is everywhere. This resonates poorly with the typical model design, which usually has one or a few noise sources (see Figure 4). Again, it should be pointed out that this does not necessarily mean that such models are inaccurate; it is entirely possible that one noise source can mimic the aggregate effect of several sources.

Figure 4: An archetypical signal detection model. An auditory model transforms the acoustic signal $X$ into a sensory pattern $R$, which is observed by the decision device and leads to a decision $D$. Noise can be inserted at any point in the system, but most commonly only at one or a few points.

1.3 Estimation of psychophysical thresholds

Perhaps the most common goal in psychophysical experiments is to estimate the threshold parameter of the PF, or the closely related just noticeable difference (JND). As previously noted, the threshold (or JND) corresponds to a particular point on the PF, which the experimenters are free to set at will. However, the point on the PF that can be most accurately estimated is the point of maximum slope. This is because the experiment measures the probability of detection, i.e. the ordinate of the PF, which means that the error along the abscissa, i.e. the threshold estimate, is approximately inversely proportional to the derivative of the PF. This can be expressed mathematically by the following relation:

$$\Delta x \approx \frac{dx}{d\Psi} \Delta \Psi,$$

where $\Delta x$ and $\Delta \Psi$ denotes the error ranges in the respective variables. This effect is illustrated in Figure 5. As it happens, the point of maximum slope is also the point of symmetry ($x = m$ in Eq. 1) for most forms of the PF.
Therefore it can be concluded that threshold estimation is most efficiently done by estimating $m$.

Figure 5: Error propagation in threshold estimation. A fixed error in $\Psi(x)$ results in a variable error in $x$, depending on the slope of the psychometric function. To minimize the error in $x$, the experiment should be designed to find the point of maximum slope.

An experimental threshold estimation procedure generally consists of two steps:

1. **Choose stimulus levels.** The simplest approach is to use a fixed number of presentations at each level, where the levels have been decided beforehand (known as the method of constant stimuli). Clearly this is quite inefficient, as only a fraction of the presentations are in the vicinity of the threshold. Since the threshold position is not known beforehand, the presentation levels in an efficient method must be chosen online, making use of the information gained so far at all times. Such methods are known as adaptive procedures. The simplest possible adaptive procedure is the up-down method, introduced by Dixon and Mood in 1948 [20], in which the level is decreased by one step for each positive response and increased one step for each negative response. As such, the method will converge to the 50% point of the PF, which may not be appropriate, especially if the parameter $\gamma$ is large. Levitt extended the method so that more than one successive correct response is needed to decrease the level, which enables the method to converge to other probabilities [51]. Still, these methods suffer from using a fixed step size, which is wasteful; it is more efficient to use a large step size initially to find the approximate
threshold position, and then use a smaller step size to pinpoint the position more accurately. The PEST method [88] uses this approach, but it is based on completely heuristic rules. In order to apply formal sequential statistics, it is necessary to involve the PF, usually in parametric form. This also gives the opportunity to estimate other parameters than the threshold, most commonly the slope. An early such procedure was the sequential maximum likelihood (ML) method introduced by Hall [36]. Following this, several methods have been proposed that make use of ML or Bayesian estimates; the latter are reviewed in Paper C. It should be noted that these methods do not actively optimize the presentation level according to a formal criterion; it is simply assumed that the optimal choice of level for the next presentation is identical to the current threshold estimate. While this is true under certain conditions, it is not true in general (although it may be a good approximation in many practical cases). Depending on the form of the PF, the probability distribution for the threshold, and the optimization criterion, the optimal stimulus level may deviate from the threshold estimate.

To find the optimal presentation level, it is necessary to “look ahead” and consider all possible outcomes of the next trial. With this in mind, it is obvious that one could look more than one step ahead. Such methods are called “non-myopic” (i.e. far-sighted), which has been suggested by several researchers, but is mostly of academic interest, as potential gains are generally quite small at large computational cost. If one is also interested in estimating the slope of the PF, it may be advantageous to randomize the presentation level slightly, to spread the measurement points evenly over the slope portion of the PF [86].

2. **Estimate threshold from responses.** This can be done either sequentially (online), updating the previous estimate after each new response, or offline, making a single estimate using all the collected data. The Bayesian approach is particularly suited for online estimation, as the expression for the posterior probability distribution easily factors into a recursive update formula, which is the basis for the algorithm presented in Paper C. One of the most common offline methods is probit analysis [93], which linearizes the PF and finds its parameters using linear regression. This type of method is mainly useful for data collected using the method of constant stimuli, as several responses are needed at each level. More recently, Kuss et al applied Monte Carlo sampling to estimate PF parameters and the corresponding confidence intervals using Bayesian inference [48]. This type of method approaches the theoretical limit of estimation accuracy; however, the estimates may still be biased, due to serial dependencies between sam-
2 Models of normal, impaired, and aided hearing

The term *Auditory model* can either refer to a model of some set of biological processes related to hearing, such as cochlear mechanics and nerve cell activity, or to models that only mimic the function of hearing, without claiming any correspondence to the biological structure. Here, we will only be concerned with models that account for the entire hearing system; more specifically, models that give an estimate of the detectability (or discrimination ability) between stimuli. Such models are usually of the functional type, as a biologically accurate model of auditory signal detection would need to be incredibly complex. In the following sections, the current state of models of normal hearing, sensorineural hearing loss and hearing with cochlear implants is discussed. It is assumed that the reader has a rudimentary understanding of the cochlear structure; if that is not the case, a good introduction is found in [58].

2.1 Models of normal hearing

The auditory filter

A useful abstraction in auditory modelling is that of the *auditory filter*, which describes the frequency response of a hypothetical auditory channel that corresponds to a specific place on the basilar membrane (BM). An array of such channels make up an auditory filterbank, that covers all audible frequencies. This abstraction has turned out to be quite relevant, as the frequency separation performed in the cochlea is maintained throughout the hearing system (known as *tonotopic* organization), from cochlear neurons to the top-level processing in the auditory cortex [87]. This view is also supported by auditory perception, as pure tones of varying frequency give rise to very different sensations. Auditory filters are specified either in the time-domain, e.g. in the form of difference equations, or in the spectral domain, usually in the form of the magnitude response. The time domain approach allows for more accurate representation of the temporal fine structure of signals, which may influence the perception of speech to some degree [73]. For the spectral domain methods, it is assumed that the signal is stationary over short periods of time in the range of 10-30 ms, so that the signal can be divided into signal blocks of such lengths and each block can be analyzed individually using e.g. the short time fourier transform (STFT). For speech analysis, this is a generally accepted assumption.
The main parameters of an auditory filter is its center frequency and bandwidth, and the relation between the two has been measured in various psychoacoustic experiments. Depending on the experimental setup, the bandwidth estimates will vary. The most commonly used measures are the critical band (CB) [98] and the equivalent rectangular bandwidth (ERB) [27]. Although similar, the CB is generally larger than the ERB, especially at low frequencies. The actual shape of the magnitude response is not clear; several functions have been proposed to describe the response. The most commonly used are the gammatone filter [63] and the roex (rounded exponential) filter [62]. The situation is complicated by evidence of nonlinear processing in the cochlea, causing the filters’ frequency responses to change with sound level. This is believed to be due to the outer hair cells (OHC) of the cochlea, which are known to actively amplify the BM vibration in a feedback loop that provides amplification and sharp frequency tuning [71,97]. The OHC provide an additive gain with limited amplitude, whose relative contribution diminishes as the input sound level increases. Therefore, the gain and frequency selectivity is greatest at low sound levels. Baker and Rosen adapted the parameters of a roex filter to match the results notched noise measurements in normal hearing listeners [6]. Their model has been adapted into the auditory model used in Paper A, and its frequency responses are shown in Figure 6.

Irino and Patterson [44] modified the gammatone filter to include level dependence, calling it the gammachirp filter. Its frequency response is similar to those in Figure 6. The gammatone and gammachirp filters are widely used, partly because they can be efficiently implemented in the time domain using recursive (IIR) filters. However, they are not without controversy: they are not able to explain the masking of Schroeder-phase complexes [47]. The roex filter is only specified in the frequency domain, thus excluded from such scrutiny.

Models of detection

As discussed in Section 1.2, there is no clear cut boundary between the “signal processing” and “decision” components in a model of perception. However, a very convenient structure is to make the decision stage as simple as possible. This is commonly done by designing the signal processing stage so that its output, the internal representation, is such that the Euclidean distance between two stimuli is directly related to the probability of detection (or the detectability $d'$). This can be seen as a multi-dimensional extension of Fechner’s perceptual scales, where the just noticeable difference corresponds to a fixed interval. If this is the case, an “internal” noise with fixed variance can be added to represent the uncertainty of hearing. This approach is used in Dau’s auditory model [17] and the related Oldenburg model [19]. This begs the question, which signal representation has this
property? The answer is found in psychoacoustic experiments. In the case of spectral discrimination, several researchers have found that the output levels of an auditory filterbank expressed in dB has this property. Florentine and Buus [24] found that a similar model describes discrimination of masked tones well. Plomp [67] found that the same was true for complex tones, although it was noted that the L1 norm fit the data as well as the Euclidean (L2) norm.

No auditory model presented to this date has claimed to be universal, in the sense that it could predict the outcome of any psychoacoustic test. It seems likely that no single representation can reflect all the properties of auditory perception. Instead, such a model would need to rely on more flexible structures such as blackboard systems, which have been used for auditory scene analysis models [21].

Figure 6: Frequency response of roex auditory filters, adapted from [6], at three different center frequencies (300, 1000, 3000 Hz) and at three different input sound levels, noted next to each curve (in dB SPL).
2.2 Models of impaired hearing

The most common form of impairment is called sensorineural hearing loss, which is thought to be intimately connected to loss of hair cells in the cochlea. Since cochlear hair cells do not regenerate, such damage is considered permanent and irreversible\(^2\). There are three main symptoms of such damage:

1. elevated hearing threshold
2. reduced dynamic range
3. reduced frequency selectivity

These individual properties of hearing are most accurately estimated in psychoacoustic experiments, although some physical measurements can also be made, such as oto-acoustic emissions and the auditory brainstem response (ABR). The hearing threshold is measured using standard pure tone audiometry, which is the most common and quite often the only diagnostic evaluation of hearing loss. An elevated hearing threshold is the most immediate deficiency of hearing loss, as many everyday sounds may be below the threshold and therefore not audible, but also the easiest to remedy (by simply amplifying the sound). But even when the sound level is adequate at all frequencies (e.g. by using a hearing aid), the speech recognition ability is generally not on par with normal hearing listeners. Most hearing impaired listeners need an increase in SNR of 3-8 dB to achieve the same speech recognition accuracy as normal hearing listeners [35]. This effect is known as the supra-threshold effects of hearing loss.

An attractive possible cause of all the above symptoms is the loss of outer hair cells (OHC). The OHC are believed to amplify low-level signals as much as 60-80 dB [71, 75], which would explain the threshold elevation for mild to moderate hearing losses. Loss of OHC can also explain reduced dynamic range, because of their nonlinear amplification properties, as discussed in the previous section. Also, experiments using ototoxic drugs that disable OHC function show that the OHC seem to cause a sharpening of frequency selectivity [39]. Thus, a very straightforward way of modelling cochlear hearing loss is to use auditory filters with a separate OHC stage and reduce its activity to match a certain impairment. In [59], Moore and Glasberg find that attributing 80% of the threshold elevation to OHC damage and 20% to IHC damage gives a good fit to loudness-matching data. The results presented in Paper A suggest that reduction of OHC gain and IHC activity only explains a small portion of the supra-threshold deficit

\(^2\)Recent advances in stem cell therapy suggest that hair cell regeneration may be possible. However, clinical applications are not to be expected for many years, if not decades [55].
for speech recognition in noise. However, the remaining deficit could be explained by cognitive factors or shortcomings in the methodology.

In more severe cases of hearing loss, there is evidence supporting the existence of “dead” regions, i.e. portions of the BM that gives no response regardless of sound level. Due to the overlapping nature of the auditory filters, these regions may be difficult to detect without using invasive methods. Several psychoacoustic tests have been developed, where various forms of masking is used to to cancel the overlap effect, although competing measures may give conflicting results [84].

It is interesting to note that subjects with cochlear hearing loss seem to have an intensity discrimination ability that is roughly equal to that of normal hearing listeners [25]. This implies that a model of impaired hearing do not need to have an increased amount of noise added to the signal and decision stages.

2.3 Cochlear implant models

Models of hearing with CI are generally concerned with modelling one or several of the following stages:

1. the induced electrical field in the cochlear nerve tissue as a result of electrode activation
2. the neural activity as a function of the (time-varying) electrical field in the auditory nerve cells
3. the subject’s signal detection abilities as a function of the neural activity

These stages show the width of knowledge necessary to be able to model hearing with CI. The first stage is mainly an electromagnetic problem, and is most commonly solved using numerical techniques such as the finite element method. Early attempts reduced the problem to one dimension [61], while more recent attempts usually encompasses a full 3D model of the cochlea [9, 23, 38, 69]. Figure 7 shows a rendition of such a model. The second stage is modelled using either measurements from animal cochleae or numerical analysis of nerve cells, based on e.g. the celebrated Hodgkin-Huxley equations [42]. Measurements from living human cochleae are not available, as it would require a very invasive procedure. Most studies are done on the cat, whose cochlea is fairly similar to the human counterpart. However, there are some differences, such as the myelination of cell bodies [80, 81]. Therefore, it is not clear how well such data can represent the

\(^3\)Any model of CI hearing is not complete without a model of the signal processing in the device. This will not be covered here, since the processing varies between vendors and the exact methods are proprietary. The basic strategies, however, are fairly simple. For a review, see [94].
Figure 7: A 3-dimensional computer model of a cochlea. The spread of current can be estimated using the finite element method.

human cochlea. The computational models may or may not be more accurate than the animal models, as there is no direct way of validating such models. The third stage is rarely included in CI models, as they are usually not concerned with modelling the user’s signal detection abilities. However, there is no obvious reason why this stage should differ much from the corresponding stage in models of normal or impaired hearing, at least in the case of simple discrimination experiments. In the case of speech recognition, it has been shown that CI users can have reduced cognitive skills, which may influence their performance on speech [3, 66]. However, it is not clear how such deficits can be included in a model of signal detection.

3 Auditory information processing of speech

An overarching view offered in this thesis is that the brain is essentially an information processor that interprets and makes decisions based on the incoming sensory signals. This section discusses the application of information theory to the quantification of human (or machine) speech recognition ability. The basic premise is that recognition performance increases monotonically as a function of the available information (the exact relation is discussed later in this section). This means that maximizing recognition performance is equivalent to maximizing the information rate. If this is a
reasonable assertion, then it follows logically that the goal of aided hearing is to maximize the information available to the decision device (i.e. the brain of the listener).

**Historic background**

After Shannon introduced information theory in his seminal 1948 publication [77], a period of great enthusiasm followed; the implications of the theory seemed far-reaching and some psychologists were optimistic about the theory being able to model human information processing. Many researchers noted the analogy with speech communication, notably George Miller and Colin Cherry, who both described the set of phonemes as a discrete alphabet being transmitted over a noisy, continuous channel [11, 57]. Shannon himself had an interest in speech and language; in a cleverly designed human experiment, he derived an estimate of the entropy of English text [78]. However, he was skeptical of many applications of his theory, and rightly so; over time it has become increasingly clear that human perception, and memory especially, is too complex to be captured by simple mathematical relations. Duncan Luce, a researcher who partook in these events, has written a compelling review of this period [54].

Surprisingly, in the case of speech perception, the biggest innovation took place long before Shannon’s revolution. In the 1920’s at Bell Labs, Harvey Fletcher did pioneering studies about how signal degradations affected the intelligibility of speech. By examining the perception of high- and low-pass-filtered speech, he found that intelligibility could be computed as a nonlinear sum over the signal quality (sound level, frequency distortion and noise) in each frequency band. He named this measure the *Articulation Index* (AI), which was used to quantify the speech transmission quality of a communication channel. This measure is still in use today, although in a slightly modified form, now going by the name SII (Speech Intelligibility Index, [4]). Although the AI was developed many years before information theory was introduced, it bears a striking similarity to Shannon’s capacity of a Gaussian channel [2].

The work of Fletcher was extended by Edith Corliss, who in the 1960’s noted that “the ear operates with characteristics that approximate a communication channel”, and proceeded to quantify the properties of that channel using data from psycho-acoustic measurements [14, 15]. But at this point in time, speech and hearing were still treated separately, i.e., the entropy rate of speech was determined and then compared to the capacity of the hearing channel. Arne Leijon estimated the information rate through the whole speech chain, i.e., articulation, acoustic transmission and perception [49] [50]. This was done by simulating short speech segments using hidden Markov models (HMMs). In this way, the hearing channel’s capacity could be estimated for the case of speech-like signals, rather than arbitrary
signals as was the case in earlier studies. The method presented in this thesis is a continuation of his work.

3.1 A model of speech communication

Perhaps the most important functions of a hearing aid is to restore the wearer’s ability to communicate using speech. For this reason, a model of (one-way) speech communication has been implemented, consisting of a speaker, a medium and a listener. It is important that the communication chain is initiated with the speaker’s intended message, not only the actual speech signal. A truly general model of speech communication needs to allow for mispronunciations, stuttering, dialects, etc: situations that do occur in real life communication that can make speech recognition difficult, for humans and machines. Figure 8 illustrates a real world speech communication scenario, where each stage of the process is a component in the corresponding signal processing model below it. In the case of aided hearing, a model of the hearing instrument is easily inserted into the signal chain between the noisy signal and the auditory model. This structure is used throughout this thesis to study the process of human speech recognition.

3.2 Probabilistic speech models

What is a speech model? It may be many things depending on the application, covering anything from tongue and jaw movements to the structure of language. Here, a speech model is a representation of the range of pos-
sible articulations of a set of predefined words or sentences. While written language is deterministic, in the sense that the spelling of a word (usually) does not vary between authors, speech is intrinsically variable. No two articulations of a particular word or sentence will ever be exactly alike, even when uttered by the same speaker (as illustrated by Figure 9). To un-

Figure 9: Due to inexact muscle control and changing circumstances, articulation of a fixed word will always include variations in pronunciation. The situation is analogous to the relation between typed text and handwriting.

derstand speech, a listener must disregard certain variations in spectrum, prosody and timing; something which has been proven very difficult to do automatically. The listener should also keep in mind which variations occur frequently and which are less common. This motivates the use of a probabilistic model, that expresses the probability of a given articulation, defined as $P_{X|\theta}(x|\theta)$, where $x$ is a signal representation of the utterance and $\theta$ is the articulation model for the utterance. A big advantage of using a probabilistic speech model is that it can be used both for recognition (using e.g. MAP classification) and synthesis (by drawing samples from the distribution). This is an essential trait for the presented framework, as it enables optimal classification according to the MAP rule.

Hidden Markov models

The most commonly used probabilistic speech model is the hidden Markov model (HMM). It was originally proposed by Baum and Petrie [7] as a way of modelling Markov sequences with noisy observations. It is now widely used for automatic speech recognition [45], and to a lesser extent for speech synthesis. One of the main advantages is that HMMs can be trained efficiently using the expectation maximization (EM) algorithm [18].

Many variations of the standard HMM formulation have been proposed in order to improve their speech modelling capabilities. One limitation in standard HMMs is the distribution of time spent in each state, which follows a geometric distribution. Several researchers have proposed to include explicit duration modelling, and although it generally complicates the training procedure, it can improve the perceptual quality of synthesized speech [95]
and may also improve automatic speech recognition performance [76]. Another possible shortcoming is the large number of parameters required to represent a word accurately; in the standard formulation, each state has its own “output” probability distribution. Thus, a large set of training data is required to avoid overfitting the model the data. If little training data is available, even as little as one utterance of each word, a useful modification is the tied-mixture HMM [8]. There, a single output distribution is trained on the entire data set (most commonly a mixture model, where each component represents a phoneme or speech sound). Then, the individual word HMMs are only allowed to control the mixture weights of this distribution as their output.

It should be noted that the HMM formulation of the probability distribution of speech is necessarily approximate; real speech is not generated by a first order Markov chain, but rather by an extremely complex neural and motor process. Although attempts have been made at creating such “complete” models [31, 32], they are not yet advanced enough to perform speech synthesis or speech recognition near the state of the art level.

**Modelling of noisy speech**

As a matter of convenience, it is also suitable to use HMMs to represent noisy speech, i.e. superposed speech and background noise. Using the same procedure as for speech, an HMM can be trained on a corpus of noise data. As noise by definition does not carry any useful information, it is not critical that the temporal dynamics are modelled accurately, and it should suffice to use only one or a few states in the model. A model of noisy speech can then be created by combining the respective speech and noise models, where the new model will have one state for every combination of speech state and noise state and the respective output distributions are combined.

**Speech features**

The HMM structure is “feature-agnostic”; it is simply a probability distribution and does not care about the meaning of the numbers it represents. One is therefore free to transform the speech signal into (or extract) any set of features that describe the signal. The goal of feature extraction is to remove as much redundancy as possible while maintaining the meaningful information. The most important feature of speech, that allows us to differentiate the information-bearing speech sounds (phonemes), is the position of the speech organ, i.e. the tongue, jaw, lips, etc. During speech, these are moving fairly slowly compared to the fast vibrations of the vocal folds. Therefore, the most common first step in speech processing is to divide the signal into short, about 20 ms, frames. For such short durations, the speech organ can be assumed to have a fixed position. Then, the state of
the speech organ is evaluated indirectly by analyzing the spectral content of each frame. A common approach is to estimate the spectrum using an auditory filterbank, known as the auditory spectrum. There are two main advantages of this approach: firstly, since humans are very good at recognizing speech, it should be advantageous to mimic human speech perception to some degree [41]. Secondly, speech is generally intended for human ears, and speakers may adapt their speaking style to facilitate human listening (known as the Lombard effect) [91]. For the purposes of this thesis, the use of auditory spectra at this stage simplifies the application of the auditory model at the recognition stage.

3.3 Estimation of speech intelligibility

The probabilistic model of speech communication presented in this thesis is ideally suited for estimation of speech intelligibility. By applying standard results from information theory, the probability of correct recognition of a word or phrase can be estimated. Mutual information (MI) is defined as the shared information between two random variables. Here, we are concerned with how much information is shared between the speaker’s intended word sequence \( W \) and the listener’s sensory pattern \( R \), as illustrated in Figure 8.

It is defined as

\[
I(W; R) = \mathbb{E} \left[ \log \left( \frac{P_{R,W}(r, w)}{P_R(r)P_W(w)} \right) \right] = \mathbb{E} \left[ \log \left( \frac{P_{R|W}(r|w)}{P_R(r)} \right) \right],
\]

i.e. the expectation of the log ratio between the conditional probability and the marginal probability. There is no closed-form solution for this expression when \( R \) is represented by a mixture model or an HMM, as is the case in the presented framework. However, it can be estimated with arbitrary precision using stochastic integration. To this end, random word sequences \( w_i \) are generated, which are passed through the speech and hearing models to generate the corresponding sensory pattern sequence \( r_i \). The MI is estimated by averaging over \( N \) such terms,

\[
I(W; R) \approx \hat{I}(W; R) = \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{P_{R|W}(r_i|w_i)}{P_R(r_i)} \right).
\]

This estimate will converge to the true MI of the model as \( N \) goes to infinity. However, the model itself may give rise to bias in the MI estimate, if the true distribution is on a manifold or if the model has been overfitted to data [60].

The relation between the MI estimate and the recognition performance can be derived using rate-distortion theory. The Blahut algorithm (see e.g. [16]) can be used to compute an upper bound on the recognition score as a function of the MI per word, given an error measure and a probability
distribution for the set of available words. The error measure used here is the number of incorrectly recognized words divided by the total number of words, giving an error between 0 and 1. This can be seen as a normalized version of the Hamming distance [37]. This error measure is limited to the case where the original and recognized sequences has the same number of words\textsuperscript{4}. The probability distribution for the words will affect the average recognition performance, as seen in Figure 10. The word frequencies in

![Mutual Information vs Recognition Score](image)

**Figure 10:** The information needed for a particular recognition score depends on the prior probability distribution of the available symbols to be recognized.

the English language approximately follow a “Zipf” distribution [52], and it can be assumed that the distribution will be similar for other languages. In many experimental methods such as Hagerman’s sentences [34], a small set of words are presented with uniform probability. As seen in the figure, these situations will give rise to different recognition performance.

\textsuperscript{4}This may not always be the case for speech recognition, as two words may be perceived as one or vice versa. A more general error measure is the Levenshtein distance (also known as edit distance), which counts the number of edits, i.e. insertions, deletions or substitutions, between the original and the recognized sequence. However, this is less relevant here, as the sentence recognition experiment that is used, Hagerman’s sentences, has a fixed number of words in each sentence.
Rationale for the MI method

The presented MI approach may seem convoluted, when recognition performance can be estimated directly, by simply simulating a test scenario and counting the number of correct responses. However, the MI approach is in many cases more efficient, as each iteration gives an MI estimate which is a continuous number, thus getting better resolution than only looking at the simulated response, which is binary. This effect is illustrated in Figure 11. Please note that this is only a tendency; there may be cases where the reverse is true. Also, the MI measure is less dependent of the complexity of

![Figure 11: Comparison of the mean square error in estimation of recognition probability, using the MI method and direct simulation. The recognition problem is discrimination of two Gaussian random variables with unity variance but different means. The MI method gives much smaller errors after only a few iterations, but the two methods give similar errors for large numbers of iterations.](image-url)

the speech material (except that it cannot exceed the information rate of the speech source itself), which allows prediction of recognition performance for several different tasks at once.
Competing methods

As previously discussed, Fletcher’s Articulation Index is still in use today, in the form of the SII. It is calculated as a weighted sum of the SNR in a number of frequency bands in the range 300-8000 Hz, as frequencies outside this range contribute very little to the intelligibility. The index is defined within the range 0 to 1, which corresponds to completely unintelligible and fully intelligible, respectively. A value of 0.5 does not correspond to a situation where a listener will hear 50% of the words spoken; the relation between SII and recognition score will depend on the complexity of the speech material, but is always a monotonically increasing function. The SII contains a correction for elevated thresholds, but does not correct for the supra-threshold deficits of hearing loss. However, several researchers have suggested extensions that account for this effect, e.g. Pavlovic et al. [64], Ching et al. [12], and Rankovic [1]. The latter included corrections for dead cochlear regions, but found this to give no improvement in predictive power. Pavlovic’s method is used as a reference in Paper A. Modifications to the SII also exist to account for fluctuating or modulated background noise [70].

The SII was extended by Steeneken and Houtgast to correct for non-additive distortions to the speech signal, such as reverberation and clipping [82]. Their measure is known as the Speech Transmission Index (STI). It is computed similarly to the SII, but the SNR in each frequency band is derived from the modulation transfer function (MTF) of the transmission channel. The authors claim that the STI estimate is within 1 dB SNR of the true intelligibility.

Future outlook

The idea of applying formal information theoretic methods to predict speech intelligibility has been around for many decades, but it has not been widely adopted. One of the main reasons for this is that a complete analysis remains intractable; the curse of dimensionality tells us that we need an enormous amount of data to accurately model the probability distribution of spoken words. But as databases grow along with computing power, the models will get more and more accurate. It seems plausible that these methods will gain popularity, making ad hoc methods lose relevance. In the simulations in Paper A, there is a fairly good correspondence with the normal hearing data, while the model fails to predict the effects of hearing impairment. It is conjectured that prediction is still possible using this approach, but each stage of the model needs to be improved.

4 Summary of Contributions

Two main contributions have been presented:
• A novel framework for estimating speech recognition ability using individualized auditory models (Paper A and B).

• A new method for finding multiple psychophysical thresholds efficiently utilizing known statistical dependencies between the thresholds (Paper C).

This work is described in more detail in three research papers that are included in the thesis. A short summary of each paper is presented below.

**Paper A: An Information Theoretic Approach to Predict Speech Intelligibility for Listeners with Normal and Impaired Hearing**

This paper presents the foundation for the information rate approach for predicting individual speech recognition abilities. Tied-mixture HMMs are used to represent the words of Hagerman’s sentences mixed with speech weighted noise. The noisy speech signals are passed through an auditory model that contains OHC and IHC stages. For hearing impaired listeners, the functionality of these components are modified to match the individual hearing thresholds of each subject. The model is used to estimate the information rate available to each listener. Using rate-distortion theory, the average word recognition rate can be calculated from the information rate. This result is then compared to experimental results for both normal hearing and hearing impaired listeners. The simulation of normal hearing listeners gives results that match the experimental data well. However, for hearing impaired listeners, the method underestimates their deficits greatly. This indicates that there are other supra-threshold effects of hearing loss that are not captured by this auditory model. These effects can be due to both cognitive and signal transmission deficits. The discrepancy may also be due to inadequate speech and language models.

**Paper B: Prediction of speech recognition in cochlear implant users by adapting auditory models to psychophysical data**

Here, the basic approach from Paper A is adapted to predict speech recognition for CI users. Two models of hearing with CI have been developed: a simple “functional signal processing model” and a complex model that simulates current spread in the cochlea and electric activity in the auditory nerve. Both models are adapted to individual users by simulating a psychoacoustic spectral discrimination task and setting the model parameters to match the user’s experimental results as closely as possible. The adapted models are then used to predict individual speech recognition thresholds using the Hagerman task. It is found that the predictions are significantly correlated with the experimental results for both CI models. The presented
framework may be used to optimize and evaluate CI encoding strategies in the future.

**Paper C: Bayesian Optimal Pure Tone Audiometry with Prior Knowledge**

The Beltone database, which contains over 100,000 audiograms, is used to create a probability distribution for the hearing thresholds of hearing impaired persons. The distribution is represented by a GMM with 10 components. This serves as the prior distribution for an optimal Bayesian experiment method, designed for performing pure toneaudiometry using as few tone presentations as possible. A Bayesian update method has been designed to update the distribution after receiving a yes/no response from a listener. Presentation parameters (tone frequency and sound level) are optimized before each presentation by maximizing the expected utility after the presentation. Several utility functions (or objective functions) have been implemented, including the entropy function. Performance is evaluated using simulated test persons, with hearing thresholds drawn randomly from the prior distribution. It was found that an average of 48 presentations was needed to achieve the same average error as the standard method, which requires 135 presentations on average.

**5 Conclusions**

In the course of the work for this thesis, two separate goals have been pursued. The first goal has been to implement a framework that can simulate speech communication, including models of normal hearing, impaired hearing and hearing with cochlear implants. The hearing models have been designed so as to be able to adapt its properties to a specific individual. The framework has been tested using the speech test Hagerman’s sentences and compared to clinical measurements. The results showed a fairly good match for the CI and normal hearing data, but the model of cochlear hearing impairment greatly overestimated the subjects’ speech recognition abilities. The second goal was to design a method for minimizing the number of tone presentations needed for a clinical hearing threshold assessment. A statistical model of hearing thresholds with a prior distribution approximated from a large database of clinically measured thresholds was designed. Each tone presentation was chosen according to an optimality criterion, and a Bayesian update formula was derived to update the model after each response. It was found through simulation that the method greatly reduced the number of presentations needed, as compared with the standard method.
References


Introduction


