DEPARTMENT OF ELECTRONIC ENGINEERING
UNIVERSITY COLLEGE GALWAY

Development of Active Cochlear Models for Speech Processing

Brian McDonagh B.E.

Submitted to the National University of Ireland in fulfillment of the requirements for the degree of Master of Engineering Science

Head of Department
Prof. David J. Wilcox

Project Supervisor
Dr. Eliathamby Ambikairajah

October 1996
# TABLE OF CONTENTS

## Introduction

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cochlear Modelling</td>
<td></td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Lyons model</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Seneffs model</td>
<td>11</td>
</tr>
<tr>
<td>2.4 Kates model</td>
<td>12</td>
</tr>
<tr>
<td>2.5 Ambikairajahs model</td>
<td>15</td>
</tr>
<tr>
<td>2.6 Selection of frequency scale</td>
<td>18</td>
</tr>
<tr>
<td>2.7 Summary</td>
<td>26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active gammatone model for the response of auditory nerve fibers (Model 1)</td>
<td></td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>27</td>
</tr>
<tr>
<td>3.2 The narrowband filter</td>
<td>27</td>
</tr>
<tr>
<td>3.3 Feedback control of bandwidth</td>
<td>31</td>
</tr>
<tr>
<td>3.4 Travelling wave delay</td>
<td>34</td>
</tr>
<tr>
<td>3.5 Table of values for the gammatone single section model</td>
<td>34</td>
</tr>
<tr>
<td>3.6 Hair cell and synapse model</td>
<td>36</td>
</tr>
<tr>
<td>3.7 Adaptation to filterbank formation</td>
<td>37</td>
</tr>
<tr>
<td>3.8 Tuning characteristics of the filterbank at sample CF</td>
<td>38</td>
</tr>
<tr>
<td>3.9 Summary</td>
<td>42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development of alternative active models (Models 2 and 3)</td>
<td></td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>43</td>
</tr>
<tr>
<td>4.2 Parallel alternative filterbank (Model 2)</td>
<td>45</td>
</tr>
<tr>
<td>4.2.1 Basic plan for alternative single channel model</td>
<td>45</td>
</tr>
<tr>
<td>4.2.2 Digital filter design</td>
<td>46</td>
</tr>
<tr>
<td>4.2.3 Inclusion of compressive mechanism into the model</td>
<td>49</td>
</tr>
<tr>
<td>4.2.4 Final version of model 2</td>
<td>56</td>
</tr>
<tr>
<td>4.2.5 Evaluation of model 2 performance</td>
<td>58</td>
</tr>
</tbody>
</table>
4.3 Transmission line adaptation for realistic compression ratios (Model 3) 63
4.3.1 Proposal for realistic compression ratios in a cochlear model 63
4.3.2 Modified Transmission Line (MTL) Model 64
4.3.3 Evaluation of performance of MTL 68
4.4 Summary 69

5. Inner hair cell/synapse model
5.1 Introduction 70
5.2 Transduction models 73
5.3 The inner hair cell 77
5.4 The synapse model of Westerman and Smith 79
5.5 Outputs of synapse model to tones at high and low Q 82
5.6 Summary 84

6. Speech recognition results
6.1 Introduction 85
6.2 Neural networks and backpropagation 90
6.3 Training the neural network with the E-set speech database 93
6.4 Recognition results 96
6.5 Summary 99

7. Conclusions 100

References 103

Appendix A Centre Frequency Shift in Digital Filter Design 107

Appendix B Sound Pressure Level (SPL) 111

Appendix C Publication 113

v
Abstract

This thesis describes the development of active cochlear models with realistic temporal and spectral characteristics. The human auditory system can hear a vast range of sounds spanning some twelve orders of magnitude of input pressure intensity. The neurons which send sound signals to the brain must operate over a much reduced range of intensities, since there is a biological limitation with regard to the number of impulses per unit time which can be transmitted along the nerves. Thus the auditory periphery is presented with the problem of encoding up to 120 dB of possible speech levels with neurons which have a dynamic range of around 40 dB. This thesis describes three cochlear models with an adaptive, active element which simulates the compressive action of the outer hair cells of the cochlea, two of which have been developed by the author. Emphasis is placed on producing accurate temporal responses, the input level dependent nature of the cochlear filters, and the resulting neural tuning curves.

The first model is a cochlear filterbank implementation based on a single channel gammatone filter. The second is developed to reproduce similar results with regard to the criteria mentioned above, but in a much more computationally efficient manner, since the models are intended for speech processing applications. It is seen that the single channel models have inherent difficulties in producing substantial compression ratios while maintaining acceptable latencies and cochlear bandwidths. The third model which is developed is a transmission line model which contains a feedback loop to alter the filter gain depending on input sound intensity. Realistic latencies and bandwidths are maintained, but due to the cascade of filter sections it is also possible to produce large compression ratios (of up to 2.5:1), enough to compress a 100 dB input range of sounds into the 40 dB dynamic range of the inner hair cells and synapse models.

To ensure that the inclusion of the active compression mechanism doesn’t impede the performance of the model vis a vis other non-active models, the models are used as front end processors to a neural network based speech recognition system. It is seen that recognition results of up to 65 % are attained, which compare well with other physiologically based recognisers. This acts as verification of the model produced by the inclusion of realistic active elements into already existing cochlear speech processors with the aim of producing realistic fine time outputs.
Chapter 1
Introduction

The motivation behind cochlear modeling as a field of research is twofold. Firstly there is the attempts of physiologists to understand the process of human hearing, itself a contributor to the more general scientific goal of understanding human cognition. The understanding of the way in which the brain deals with sensory information is, as a science, still in its infancy. Yet this is a vital first step in what is arguably sciences greatest project, that is to understand the way the human brain works. Thus the positing of models of the human auditory system and attempted evaluations thereof is a legitimate scientific project in its own right.

In a field of research with such immediate human connections it becomes quickly apparent that the range of possible applications of even approximate models of the process of hearing will have enormous practical implications. There is an enormous amount of research being done by many of the world’s leading computer hardware and software firms to develop algorithms and techniques which are able to simulate human hearing. While truly revolutionary applications will only become available with more sophisticated theories of language comprehension, an area which goes right to the heart of Artificial Intelligence, there are already many applications becoming apparent which exploit more straightforward speech recognition algorithms. These range from Computer Voice Response systems, whereby the manually operated interface between a computer and its operating system is replaced by speech activated commands, to low-level security applications, to telecommunications applications such as automatic operator services.

As an engineering thesis, the motivation for this research is principally in the second area, i.e. that of application. As such this thesis is more concerned with performance criteria than that of physiological accuracy. The latter should not be ignored, however, since the human auditory
system is so vastly superior to any models yet developed it is reasonable to assume that by holding a models characteristics as closely as possible to the observed phenomena, the performance cannot but be enhanced.

The following is a chapter by chapter guide to this thesis.

The second chapter gives a detailed survey of some of the most influential and principal cochlear models in use. The first two models presented are popular models both of which were developed in the 1980’s (Lyons 1982, Seneff 1984). These are examined because their respective designs illustrate two major modalities of cochlear filter design. The first is an analytical technique which assumes that the operation of the cochlea can be divided into separate components, and the physiological behaviour of the cochlea is approximated using transmission line theory. The second (Seneff 1984) has a phscophysiological inspiration, in which data comes from experiments on subjects designed to get an approximation for the frequency characteristics of the cochlear membrane. The original of these two models are both passive filter based representations of the cochlea, although an implementation of a variation of Lyons model which contains an Active Gain Control mechanism (Slaney and Lyon 1983) is shown. The third model examined is that of Kates (1991, 1993) which attempts to incorporate a compressive nonlinearity to change cochlear filter shapes in response to varying input Sound Pressure Levels, and is thus concerned with similar motivations as the author in this thesis. Although Kates’ model is numerically stable, and reasonably computationally efficient, the resulting auditory tuning curves are broader than those measured physiologically. For models which are intended for use in speech processing applications it is essential that the frequency resolution of the cochlear filters is as high as possible.

This aspect of cochlear performance, along with the level dependent nature of the filter parameters, are concentrated on in the authors model, which uses as its’ foundation the cochlear modelling techniques of the Speech Research Group based in Athlone Regional Technical College, and in particular the transmission line model of Ambikairajah (1989), developed from first principles, approximating the behaviour of the membrane by means of a series of short sections with physical characteristics dependent on the position along the cochlea from the middle ear. This is a stable, efficient model with good frequency selectivity
characteristics, and is the foundation model for the authors development of a model of realistic, active cochlear mechanics.

Chapter 3 gives a detailed account of one of the models used in this research, a filterbank model based on the gammatone filter proposed by Carney (1991, 1993). In this model, a time-varying filter represents a single section of the basilar membrane, and an active feedback mechanism represents the mechanoelectrical transduction properties of the Outer Hair Cells. The model parameters are selected to try to match data from low-frequency auditory nerve fibers in cat. Although the temporal behaviour of the model is quite good, the tuning curves are rather broad, and the filter itself is quite intensive computationally. Nevertheless, this is one of the most recent active models proposed, and is used in this thesis as a benchmark by which to evaluate the performance of the models developed by the author.

Chapter 4 gives a detailed account of the development of these two other models. The first of these is a filterbank model which, although not as physiologically accurate as the gammatone model, produces similar results. From a speech processing point of view, the model is superior in that it requires far less computational effort to produces the same results, indeed the frequency selectivity of the resulting auditory tuning curves are superior to Carney's model. It will be seen that both of the two previous models have an inherent weakness as far as active cochlear compression is concerned, and that is that individual single channel sections of the cochlea are quite unable to produce substantial compression ratios. The final, and most important model of this research is a cascaded filterbank model which will be seen to be both computationally efficient and effective at simulating the nonlinear characteristics of the auditory periphery. It succeeds in compressing a 100 dB input range of sounds into the required 40 dB dynamic range of the auditory hair cells and neurons.

Chapter 5 gives an account of the various hair cell models which have been proposed. The hair cell combined with the basic cochlear model provides the overall auditory periphery model which will be used as the front end processor to the speech recognition neural network arrangement. An evaluation of the performance of the model under stimulation by cochlear inputs is carried out at this stage.
Chapter 6 gives an account of basic neural network theory, and explains the exact nature of the network to be used to verify the operation of the cochlear models developed in this research. Although the aim of the models is not exclusively aimed at speech recognition, it is a recognition task which is used first of all to compare the performance of the three models, and secondly to show that the incorporation of active cochlear mechanics does not necessarily compromise traditional performance tests. A difficult recognition task has been chosen, that of distinguishing between the spoken E-set of the alphabet (utterances of the letters B,C,D,E,G,P,T,V).

Chapter 7 presents a summary of the arguments made in the thesis and some concluding remarks.
Chapter 2
Cochlear Modelling

2.1 Introduction

Structure and Function of the Auditory Periphery

The human auditory system is a superb speech recognition system which functions to an extremely high performance level even in very noisy environments with low signal to noise ratios. An anatomical diagram of the human auditory system is given in Figure 2.1. Sound enters the pinna and is funnelled along the external auditory canal to the eardrum or tympanic membrane which converts the sound pressure vibrations into mechanical vibrations of the three bones of the middle ear, the incus, malleus and stapes. These bones perform an impedance matching function between the delicate structures of the inner ear, and the sound pressure waves of the external environment. The structure of interest as far as audition is concerned in the inner ear is the cochlea. The cochlea is a spiral shaped structure which is about 3.5 cm in length if uncoiled. A cross section through the cochlea is shown in Figure 2.2. The basilar membrane is a tough sheath which runs along the length of the cochlea, separating it into two chambers (scalae), which are continuous at the apex, or helicotrema. When the stapes vibrates against the 'oval window' the fluid in the scala vestibuli vibrates which in turn sets the membrane vibrating. Running along the length of the membrane is the ‘Organ of Corti’, which contains inner and outer hair cells. The deflections of the basilar membrane cause the cilia of these hair cells to bend, which in turn triggers the hair cells to send nerve impulses towards the brain. This mechanical to electrical transduction is an essential function of the auditory system.

Pioneering work in the study of the cochlear membrane was carried out by the physicist Georg von Bekesy (1960), who carried out experiments on the cochlea which led to his positing the model of the basilar membrane functionality shown in Figure 2.3.
Figure 2.1  Human Auditory System

Figure 2.2  Cross Section of the Cochlea
The basilar membrane is narrow and stiff at the basal end (the end nearest the stapes) and is wider and more flexible at the apical end. Membrane displacement is a travelling wave that moves along the membrane in response to input pressure variations, i.e. a sound stimulus. Different points along the membrane have their maximum movement at a particular frequency, i.e. they resonate at that frequency, after which the traveling wave decays rapidly. Von Bekesy arrived at the following empirical formula to relate distance from the basal end of the membrane to resonant frequency at that point:

$$f_p(x) = 16000 \times 10^{-0.667x} \text{Hz}$$

(2.1)

with $x$ in centimetres.

Thus the essential function of the cochlear membrane is to act as a frequency analyser of the sounds which are input to the ear, resolving an input into its constituent frequencies and sending nerve impulses for each frequency, or section of the cochlea, to the brain.

**Computational models of the cochlea**

As the body of experimental evidence regarding the structure and function of the auditory periphery grew, many researchers attempted to incorporate this knowledge into the design of systems for speech analysis, and speech and speaker recognition. This led to the development of computational auditory models, which attempt to capture the salient features of auditory behaviour, though not necessarily in a strict physiological manner. In this section some of the many computational models which have been developed will be introduced (Lyons 1982, 1983, 1984, Seneff 1984, 1985, Kates 1991, Ambikairajah 1989), and the relevance of
these models in the context of the active models developed by the author including some aspects of their performance with regard to speech processing is discussed.

2.2 Lyons Model

An important and often referenced model is that developed by Lyons (1982). In this model the basilar membrane is characterized using a cascade/parallel combination of notch filters and resonators. This model assumes that the operation of the cochlea can be divided into two separate components. Firstly the compression of the basilar membrane is implemented by means of a linear model, and a secondary, nonlinear component is used for transduction and compression. It is the basilar membrane model which is implemented as a cascade/parallel combination of notch filters and resonators, and it is based on the passive long wave transmission line model of Zweig et al. (1976). A “second filter” in the form of a complex zero (Allen, 1980) is added to each resonator. A block diagram of the basilar membrane model is given in Figure 2.3 and the pole-zero plots (in the S-domain) for one section of the model are shown in Figure 2.4. The cascade of notch filters results in a very sharp cutoff along the membrane model, due to the summing effect of each filter in the cascade.

![Block Diagram of Lyon's cochlear model](image-url)

*Figure 2.3* *Block Diagram of Lyon's cochlear model*
In order to compress the huge dynamic range of sounds that the ear can hear into a smaller and more physiologically manageable range of Sound Pressure Levels (see Appendix B for an explanation of Sound Pressure Level), a transduction mechanism can be implemented which consists of a linear half-wave rectifier plus several nested stages of coupled Automatic Gain Controls (see below). The reason for such a Coupled Gain Control mechanism is the possibility that sensory neurons can use a feedback mechanism to reduce the gain factor of themselves and neighbouring cells, so called lateral inhibition. Lyons reasoned that a coupled system was required in order to preserve local spectral contrast while de-emphasising overall loudness changes and spectral tilt, which would be difficult to achieve if a simple point compressive nonlinearity was used.

Variations on Lyons original model

There are several variations of Lyons original (1982) model. Lyon himself (1983) used a cascade/parallel filterbank as a front end processor for a system to implement binaural localisation of sound sources in which the choice of filter parameters was motivated by the phenomenon of critical bands such that channels with low centre frequency had a constant bandwidth (approx. 100 Hz), while high frequency channels had a constant Q factor. Another variation of the original cascade/parallel filterbank, which used only a single canonic second order section for each channel was proposed in Lyon (1984). This model also included a transduction mechanism with cascaded, rather than nested, AGC’s. The cascaded AGC’s modelled different time constants and spatial extents by having different parameters. The AGC’s were followed by a physiologically-based model for hair cell transduction and the auditory neuron. Although no specific sharpening mechanism was explicitly included in the
model, the resulting tuning curves of the auditory neuron model when measured (Lyon and Dyer, 1986) were sharper than the mechanical tuning curves. The model performed well as a front end processor for a vector quantisation classifier used for digit recognition (Loeb and Lyon (1987), due to the enhanced tuning curves mentioned above. These effects were concluded to be side effects of the inclusion of the AGC’s into the model. While the coupled AGC mechanism has no direct physiological correlate, Lyon and Dyer (1986) suggested that the mechanism was due to the influence of efferent nerve fibers on the OHC’s.

Another variation on the original model was proposed by Lyon and Mead (1988), in which the propagation of waves along the basilar membrane continued to be represented by a cascade of second order filters, and which retained a mechanism for automatic gain control, but which contained a number of important differences. Firstly, the one-dimensional transmission line model was replaced by a two dimensional short wave formulation, and secondly, the automatic gain control function was implemented by a negative damping term in the cochlear equation to represent the outer hair cells. This technique of time varying gain factors resulted in filter Q-factors which change in time as a function of input signal Sound Pressure Level, contrasting with the time-invariant basilar membrane filters used in the previous models. A digital implementation of this model was described by Slaney and Lyon (1983), a block diagram of which is shown in Figure 2.3. Note that the feedback loop which controls the basilar membrane filter characteristics is included in the model, and is implemented by means of a set of parallel, coupled lowpass filters, whose outputs are summed, and used to control the damping parameters of the BM second order filters.

Figure 2.5 Block diagram of Slaney and Lyon’s cochlear model
2.3 Seneffs Model

The models described in Section 2.2 are good examples of models which are derived computationally from examining the way in which sound pressure waves are physically propagated along the basilar membrane. There are alternative ways in which to choose the criteria by which the auditory model should operate. A typical example is that developed by Seneff (1984, 1985, 1986, 1988) which uses psychophysical (rather than physiological) data for the design of the cochlear filterbank.

A block diagram of the model is given in Figure 2.6. The basic arrangement looks similar to that of Lyon, with the addition of a middle ear filter (a cascade of four complex zero pairs) to remove very low and very high frequencies from the input sound stimulus.

![Block diagram of Seneff's Cochlear Model](image_url)

**Figure 2.6** Block diagram of Seneff's Cochlear Model

In Seneff's original (1984, 1985) model, the basilar membrane is represented by 32 parallel filters tuned to frequencies in the range from 200 Hz to 3.2 kHz. The filter centre frequencies and bandwidths were chosen to satisfy the critical bandwidth criterion of Zwicker (1961), with filter CF's separated by half a critical band. Each filter consists of a double complex pole pair at CF, as well as several zeros on the x-axis and on the unit circle, which are required to shape the high and low frequency shapes of the filter. The active element in Seneff's model comes...
after the basilar membrane (BM) filters. A pair of AGC’s compress the dynamic range of the input signal, as well as simulate the phenomena of rapid and short-term adaptation. However, since each gain factor cannot influence the output of adjacent cochlear filters the amount of gain is limited, as it will be seen to be for the single channel models (model 1 and model 2) developed in the following chapters. The final stage of the peripheral auditory model is a half-wave rectifier. The output of each model channel is fed into a “Generalised Synchrony Detector”, whose purpose is to enhance synchrony to the channel CF, in order to emphasise speech formants. This model is also applied to the problem of pitch detection. The model is further developed in Seneff (1986,1988), by increasing the number of channels, and covering a wider frequency range (130 Hz to 6.4 kHz). For computational efficiency the filterbank is implemented as a cascade of complex high-frequency zero pairs, with taps after each zero pair to individual tuned resonators. Each resonator consists of a double complex pole pair at CF, and a double complex zero pair one octave below CF. Other changes are made to the model for the inner hair cell/auditory nerve fiber synapse.

2.4 Kates’ Model

The active model developed by Kates (1991) is of particular interest as it is similar in terms of objectives and results to model 3 developed in this work. Kates’ model concentrates on computational efficiency and numerical stability, while incorporating a compressive nonlinearity to change the filter shapes in response to changes in sound input pressure level. It also includes the ability to reproduce changes in cochlear behaviour due to outer and inner hair cell damage. The first model (1991) had auditory tuning curves which were substantially broader than those measured physiologically in animals but a revised version (Kates 1993) produced more realistic results.

Kates follows Lyons and Mead (1988) in using a cascade of active low pass filter sections, such that the output at a particular location is the sum of the effect of all the filter sections that precede it. These active low-pass filters provide unity gain at low frequencies, gain greater
than one in the vicinity of the resonance frequency, and attenuation for higher frequencies. Although none of the individual active filter sections is highly tuned the system still attains overall sharpness due to the contributions of many sections in cascade (an effect called pseudoresonance). Also, the active system adjusts the $Q$ factor of each filter section in response to signal level in an attempt to model the action of the outer hair cells in changing the cochlear filter shapes. The characteristic equation of a single filter is given by

$$H_i(s) = \frac{1 + \left(\frac{\mu + 1/Q_i}{\omega_i}\right)(s/\omega_i) + b(\mu/Q_i)(s/\omega_i)^2}{(1 + \mu s/\omega_i)[1 + s/\omega_i Q_i + (s/\omega_i)^2]}$$

(2.2)

This filter design compromised to a certain extent the sharpness of the filters in order to get realistic latencies. The net traveling wave motion at a point on the cochlea is given by

$$G_k(z) = \prod_{i=1}^{k} H_i(z)$$

(2.3)

This output is fed into a second filter which consists of a complex zero pair over a complex pole pair, transforming the cochlear velocity output into mechanical output:

$$F(s) = \frac{1 + s/\omega_o Q_o + (s/\omega_o)^2}{1 + s/\omega_p Q_p + (s/\omega_p)^2}$$

(2.4)

Kates sets $Q_o = 2Q_p$ and $\omega_o = \omega_p / 2$ resulting in group delay at low frequencies and a peak at the characteristic frequency with a notch an octave below. The $Q$ values of both filters are frequency dependent, varying on a linear cochlear distance scale for the first one, and according to the relationship $Q_p = 1.5(1 + f)$ with $f$ in kHz for the second filter.

Kates compared the operation of the second order lowpass filter with a one dimensional nonlinear transmission line model (Deng and Geisler, 1987). The results indicate that in order to produce a computationally efficient model with realistic tuning curves a compromise between the two techniques is necessary. The low-pass filters of the cascade are replaced with filters which approximate the transfer function of a section of the one-dimensional transmission line, and which are driven by the pressure output from the previous section and terminated with the characteristic impedance of the following section. This modified transmission line model looks like that shown in Figure 2.7.
Kates follows Eysholdt and Mellert (1975) in using a transfer function having one pair each of relatively high-Q poles and zeros combined with a pair each of low-Q poles and zeroes. The Modified Transmission Line (MTL) model is like the low-pass filter version, an active system, with run time adjustments to the Q values changing the shapes of the cochlear filters. The resultant adaptive system is shown in the block diagram of Figure 2.8, in which the peak displacement over a narrow frequency region is used to control the MTL section filters and the second filters.

**Figure 2.7** Single section of the modified transmission line model of Kates

![Diagram](image)

**Figure 2.8** Single section of the modified transmission line model of Kates
As shown by Kates (1993) this model is quite successful at reproducing realistic compression ratios (of about 2.5:1 over a 100 dB input range). Time-frequency neural responses to a simple speech sample also showed realistic fine time responses across the range of filters. However the bandwidth of the cochlear filters is substantially wider than those obtained by other models such as that of Ambikairajah (1989) whose model is the foundation for this work.

2.5 Ambikairajah’s Model

The digital filter simulation of the basilar membrane developed by Ambikairajah et al (1989) is the foundation model which is adapted in this work, and so is explained here in some detail. The model is developed from first principles, taking as it’s starting point the assumption that the motion of the basilar membrane can be described as a transmission line with distributed parallel and series impedances that vary continuously with distance along the membrane. Voltage and current along the transmission line are equated with pressure and volume velocity with respect to the cochlear fluid. A simple electrical representation of a section of the basilar membrane is shown in Figure 2.9.

![Electrical model of a section of the basilar membrane](Figure 2.9)
The dynamic behaviour of the basilar membrane is approximated by cascading a number of these short sections, with the parameters $M$, $L$, $C$, and $R$ dependent on position $x$ along the membrane. To make the computational task more tractable, each section is isolated from its neighbours by considering each section to be loaded by a shunt impedance representing the input impedance of the remainder of the membrane. Approximating the shunt impedance by a simple parallel connection of an inductance $M_T$ and a resistance $R_T$, and after the application of Thevenin's theorem the isolated section in Figure 2.10 is arrived at.

\[\text{Figure 2.10  Model for single section of the basilar membrane}\]

This corresponds to the following voltage transfer function:

\[
\frac{v_o(s)}{v_{The}(s)} = \frac{\left(\frac{R_T}{sL_T} + \frac{R}{L} \right)^2 + \frac{1}{LC}}{s^3 + \left(\frac{R_T}{L_T} + \frac{R}{L} \right)^2 + \left(\frac{1}{LC} + \frac{R_T}{LL_T} \right)^2 + \frac{R_T}{LL_TC}}
\]

(2.5)

This expression can be simplified to give the following equation for the voltage, or pressure transfer function:

\[
\frac{v_o(s)}{v_i(s)} = K\frac{a}{s + a} \frac{\omega_p^2}{s^2 + B_p s + \omega_p^2} \frac{s^2 + B s + \omega_i^2}{\omega_i^2}
\]

(2.6)
s  complex frequency variable

$K$  attenuation factor

$a$  low pass pole frequency

$\omega_p$  resonant pole frequency

$B_p$  pole bandwidth

$\omega_z$  resonant zero frequency with $\omega_z > \omega_p$

$B_z$  zero bandwidth

Thus there are three aspects to the filtering action of a particular section. Firstly, frequencies below the pole frequency $\omega_p$ will be attenuated by factor $K$. Secondly, frequencies close to the pole will be enhanced; and thirdly, frequencies above the zero frequency $\omega_z$ will be sharply attenuated.

The digital filter representation of a single section of the membrane is given in Figure 2.11.

**Figure 2.11**  Ambikairajah’s digital filter implementation of a basilar membrane section

128 of these filters are arranged in cascade, each tuned to a particular frequency of the auditory spectrum according to one of the frequency scales discussed in section 2.6. When an input speech stimulus is applied to the first filter the cascade should simulate the travelling wave motion of the basilar membrane.
2.6 Selection of frequency scale

For all cochlear models a pertinent issue is the exact choice of centre frequency values for the individual filter sections, and how the frequency values change along the basilar membrane. Estimates of the bandwidths and centre frequencies can be based on either direct physiological measurements, or psychoacoustic experiments in which the threshold of detectability of a particular tone is measured in the presence of competing masking tones. These different approaches produce slightly different estimates of auditory tuning characteristics. Also as measuring and experimental techniques have improved more refined estimates of the frequency distribution of the cochlea have emerged. There are several different frequency scales which have been devised and any one of them can be used to choose the centre frequencies for the filters.

The choice of frequency scale can have a significant effect on aspects of the performance of the model, for example in a speech recognition system using a cochlear model as a front end processor recognition accuracy can be affected by the distribution of centre frequencies. For example, if the utterances to be distinguished between contain large amounts of high frequency information (plosives, fricatives, etc.), then it is better to concentrate the centre frequencies more towards the high-frequency end of the auditory spectrum in order to increase the high frequency resolution. The most important scales developed thus far are:

- Bekesy scale  (Bekesy, 1960)
- ModLog scale  (Ambikairajah, 1989)
- Mel scale     (Parsons 1987)
- Bark scale    (Zwicker and Terhardt 1980)
- ERB 1         (Moore and Glasberg 1983)
- ERB 2         (Hermes and van Gestel 1991)

There follows a description of the various scales, and a discussion of their relative merits.
2.6.1 Bekesy scale

Von Bekesy deduced the original frequency scale when performing his pioneering experiments on post mortem cochleae. In 1960 he arrived at an empirical formula for relating frequency tuning of the cochlea $f_p(x)$ to distance from the basal end of the membrane:

$$f_p(x) = 16000*10^{-0.667x} \text{ Hz } \quad [0 < x < 3.5 \text{cm}]$$  \hspace{1cm} (2.7)

Substituting in the values for the frequency range

<table>
<thead>
<tr>
<th>$f_p(x)$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>3.4918</td>
</tr>
<tr>
<td>3550</td>
<td>0.9803</td>
</tr>
</tbody>
</table>

Table 2.1 Bekesy Frequency Scale

This gives the following pair of equations for the centre frequency $f_p(x_i)$ of the $i$th filter given the sectional length of $\Delta x = \frac{3.4918 - 0.9803}{127} = 0.01978$

$$x_i = 0.9803 + i*0.01978$$

$$f_{pi}(x_i) = 16000*10^{-0.667x} \text{ Hz }$$  \hspace{1cm} (2.8)

2.6.2 Mod-Log scale

Von Bekesy's scale was a good first approximation to the distribution of frequencies along the membrane, but was subject to the inaccuracies of being an empirical formula and also having been deduced by measurements on dead cochleae. It was logical to try to deduce formulae which related frequency tuning to the mechanical properties of the membrane. As stated in section 2.1, the width of the basilar membrane increases towards its apical end. An empirical formula relating basilar membrane width to distance from the basal end is given by:

$$b(x) = 0.019 + 0.0093x \quad [0 < x < 3.5 \text{cm}]$$  \hspace{1cm} (2.9)
Ambikairajah’s Mod-Log scale (1989) assumes that each sectional length $\Delta x$ varies in proportion to the width of the basilar membrane. Then the Mod-Log equations are derived as follows (Jones 1994).

$$\Delta x = k \Delta y b(x)$$

$$\frac{dy}{dx} = \frac{1}{k(0.019 + 0.0093x)}$$

$$y(x) = k_1 \ln(1 + 0.4895x) + k_2 \quad (2.10)$$

From Table 2.1, at the basal end of the membrane (filter $y = 0$) $f_p(0.9803) = 3550Hz$ and at the apical end of the membrane (filter $y = 127$) $f_p(3.4918) = 3550Hz$. Substitution into equation 2.10 yields:

$$0 = k_1 \ln(1 + 0.4895 * 0.9803) + k_2$$

$$127 = k_1 \ln(1 + 0.4895 * 3.4918) + k_2 \quad (2.11)$$

Solving gives:

$$k_1 = 209.99$$

$$k_2 = -82.29$$

Thus, from equation 2.10:

$$y(x) = 209.99 \ln(1 + 0.4895x) - 82.29 \quad (2.12)$$

The Mod-Log equations are therefore

$$x_i = \frac{y_i + 82.29}{209.99} - 1$$

$$f_p(x_i) = 16000 * 10^{-0.667x} \ Hz$$
2.6.3 The Mel scale

The third scale examined borrows terminology from music perception, in which the relative distance between two tones is expressed in a musical interval such as the semitone and octave. Pitch is defined relative to a reference frequency, and the difference or relative distance in pitch, $D$, between two notes $f_1$ and $f_2$ is given by

$$D = 1200 \cdot \log_2 \left( \frac{f_1}{f_2} \right)$$

(2.14)

The unit of pitch is the Mel, and in psychoacoustics the Mel scale is based on a subjective measure of pitch magnitude. The Mel scale is approximated by a function (Parsons 1987) which is approximately linear below 1 kHz and logarithmic above 1 kHz. The function is given by:

$$f_m = \frac{1000}{\log_{10} 2} \log_{10} \left( 1 + \frac{f}{1000} \right)$$

(2.15)

where $f$ is in Hz and $f_m$ is in mels. Substituting in the values for the frequency range:

<table>
<thead>
<tr>
<th>$f$</th>
<th>$f_m$ (mels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>104.337</td>
</tr>
<tr>
<td>3550</td>
<td>2185.867</td>
</tr>
</tbody>
</table>

Table 2.2 Mel Frequency Scale

This gives the following pair of equations for the centre frequency $f_{pi}$ of the $i$th filter, with a Mel scale interval of $\Delta f_m = \frac{2185.867 - 104.337}{127} = 16.390$:

$$f_{mi} = 2185.867 - i \cdot 16.390$$

$$f_{pi} (f_m) = 1000 \left( \log_{10} \frac{f_m}{1000} \right)$$

(2.16)
The fourth scale examined is called the Bark scale, and is another empirically based scale. In this case it is based upon measurements of the critical bandwidth, or the frequency selectivity of the human auditory system. The scale is described in detail by Zwicker and Terhardt (1980), in which the frequency response of the auditory system is seen as a bandpass filter which corresponds approximately to the tuning curves of auditory neurons. These filters are called critical band filters, and their shapes have been determined in experiments using broadband lowpass or highpass noise to mask tones. The filters are nearly symmetric on a linear frequency scale, with very sharp roll-offs.

The Bark scale is approximately linear below 500 Hz and approximately logarithmic above 500 Hz. For low frequencies the critical bandwidth remains approximately constant around 100 Hz and for higher frequencies the critical bandwidth increases with frequency. The following analytical expression is derived for mapping from the frequency domain into the critical-band rate domain:

\[
 f_B = 13 \tan^{-1}(0.76f) + 3.5 \tan^{-1}\left(\frac{f}{7.5}\right)^2
\]

where \(f_B\) is in Barks, \(f\) is in kHz and \(\tan^{-1}\) is in radians. This equation cannot be solved for \(f\) in terms of \(f_B\) and so an approximation is given in Zwicker and Terhardt (1980):

\[
 f_B = 13 \tan^{-1}(0.76f) \quad [f < 1.5 kHz]
 f_B = 8.7 + 14.2 \log_{10} f \quad [f > 1.5 kHz]
\]

Substituting in the values for the frequency range:

<table>
<thead>
<tr>
<th>(f) (kHz)</th>
<th>(f_B(f))</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>0.741</td>
</tr>
<tr>
<td>3550</td>
<td>16.577</td>
</tr>
</tbody>
</table>

*Table 2.3*  Bark Frequency Scale
This gives the following equation for the centre frequency $f_{ci}$ of the $i$th filter, with a Bark scale interval of $\Delta f_B = \frac{16.577 - 0.741}{127} = 0.1247$:

$$f_{ci} = 16.577 - i \times 0.1247$$

$$f_{ci} = 1000 \left( \frac{\tan \left( \frac{f_{ci}}{13} \right)}{0.76} \right) \quad \text{if } i < 5$$

$$f_{ci} = 1000 \times 10^{\left( \frac{f_{ci} - 8.7}{14.2} \right)} \quad \text{if } i \geq 5$$

(2.18)

2.6.5 The ERB1 scale

The original empirical anatomical findings of von Bekesy have been updated by more accurate work recently (Moore and Glasberg 1983), and indicate that the Bark scale may need to be revised. Research has shown that the equivalent rectangular bandwidth of the auditory filter continues to decrease below 500 Hz. If the critical bandwidth is related to the equivalent rectangular bandwidth of the auditory filter, this may indicate a need to revise the classical critical band function. The ERB-rate is given by:

$$E = 11.17 \ln \left| \frac{f + 0.312}{f + 14.68} \right| + 43.0$$

(2.19)

where $f$ is in kHz. Substituting in the values for the frequency range:

<table>
<thead>
<tr>
<th>$f$</th>
<th>$E(f)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>2.331</td>
</tr>
<tr>
<td>3550</td>
<td>25.665</td>
</tr>
</tbody>
</table>

Table 2.4  ERB1 Frequency Scale
This gives the following equation for the centre frequency $f_{pi}$ of the $i$th filter, with an ERB scale interval of $\Delta E = \frac{25.665 - 2.331}{127} = 0.1837$:

\[
E_i = 25.665 - i \times 0.1837
\]

(2.20)

\[
f_{pi}(E_i) = 1000 \left( \frac{14.68e^{\frac{(E_i - 0.43)}{11.17}} - 0.312}{1 - e^{\frac{(n - 0.43)}{11.17}}} \right)
\]

(2.21)

### 2.6.6 ERB scale 2

Further experiments were carried out by Hermes and van Gestel (1991) to determine whether pitch movements in speech intonation are perceived on a linear frequency scale, on a logarithmic frequency scale, or on a psychoacoustic scale representing the frequency selectivity of the auditory system. They came down in favour of the latter and proposed a psychoacoustic scale very similar to the ERB scale of Moore and Glasberg (1983). The ERB-rate expression is

\[
E = 16.7 \log_{10} \left( 1 + \frac{f}{165.4} \right)
\]

(2.22)

where $f$ is the frequency in Hz and $E$ is the ERB-rate in ERB. Substituting in the values for the frequency range:

<table>
<thead>
<tr>
<th>$f$</th>
<th>$E(f)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>2.712</td>
</tr>
<tr>
<td>3550</td>
<td>22.570</td>
</tr>
</tbody>
</table>

*Table 2.5 ERB2 Frequency Scale*

This gives the following equations for the centre frequency $f_{pi}$ of the $i$th filter with an ERB scale interval of $\Delta E = \frac{22.570 - 2.712}{127} = 0.1564$.
Figure 2.14 shows a plot of the pole frequency versus filter number for each of the frequency scales examined here.

\[ E_i = 25.570 - i * 0.1564 \]
\[ f_p(E_i) = 165.4\left(10e^{0.06E_i} - 1\right) \]  

(2.23)

2.6.7 Comparison of scales

Figure 2.14 shows a plot of the pole frequency versus filter number for each of the frequency scales examined here.

![Comparison of available cochlear frequency scales](image)

**Figure 2.14** Comparison of available cochlear frequency scales

A detailed comparison of the performance of the various scales in response to a straightforward input stimulus consisting of 5 tones is carried out by Friel (1995). The results of this analysis match the expectations from Figure 2.14 above. In particular the Bekesy scale is concentrated on the lower frequency ranges, and the low frequency resolution is very good. However there a smaller number of filters spread out over the higher frequencies indicating a poor high frequency resolution. The opposite is the case for the Mel frequency scale, in which
there are fewer frequencies covering the low frequency ranges and a large number of filters covering the high frequency ranges. This is born out by Friels (1995) examination of the pressure envelopes and frequency responses of the filter sections. The Bark scale is similar to the Mel scale and also has a bias towards the high frequency end of the auditory spectrum. The remaining three scales, the Mod-Log scale and the two ERB scales are all similar in that they are intermediate frequency scales, between those of the Bekesy and Mel scales. These scales distribute the filter sections more evenly over the entire auditory spectrum. The resulting displacement envelopes and pressure responses are more evenly spread out over the entire range of the basilar membrane. Thus if the input sound stimuli of interest are not for some reason biased towards either the low or high end of the frequency spectrum it is reasonable to assume that the best results on average will be obtained by one of the latter three models. The Mod-Log frequency scale is used from this point on in this work.

2.7 Summary

This chapter has presented a description of the human auditory periphery, particularly with regard to the action of the cochlear mechanics in converting input sound waves into a frequency transform which is sent to the brain by means of neurons attached along the length of the basilar membrane. An introduction to cochlear modelling is offered, followed by a detailed description of four relevant models. These consist of the basic model of Lyons (1982), passive in its basic form but with a possible active feedback mechanism, the active model of Seneff(1983), the cascaded, active model of Kates (1993) and the passive transmission line model of Ambikairajah (1989). Finally a comparison of the various available frequency scales was given. The next chapter presents a detailed description of an active single channel model based on a gammatone filter, and measures its temporal and spectral performance.
Chapter 3

Active gammatone model for the response of auditory nerve fibres (Model 1)

3.1 Introduction

This chapter gives a detailed account of the computational model developed by Carney to model the responses of auditory nerve fibres in cat, using a compressive nonlinearity to model the active mechanics of the basilar membrane. This model is used later on as a benchmark by which to compare the performance of the two alternative models developed by the author in this work. The inner hair cell model used in this chapter is explained in more detail in chapter 5.

This model consists of a time-varying filter, which represents a single section of the basilar membrane, with an active feedback mechanism representing the mechanoelectrical transduction properties of the Outer Hair Cells. This is followed by a model for the Inner Hair Cell, and the Inner Hair Cell-Auditory Nerve Fibre synapse. Finally the single section model is arranged in parallel filterbank formation to represent the entire cochlea.

3.2 The narrowband filter

The most interesting aspect of Carney's model is the form of the narrowband filter which is used to characterise the tuning of the basilar membrane. The two main aims of this filter are to allow continuous variation of the filter properties as a function of time and to allow arbitrary input waveforms. The reason for the time-varying nature of the filter properties is that it is desired to simulate the relatively sharp tuning of the membrane at low input Sound Pressure Levels, and to allow for a broadening of the membrane response curve at higher input
amplitudes. Arbitrary input stimuli are accommodated by developing the relevant difference equations to allow any digital stimulus to be input to the model (in this work the inputs consist of pure tones during testing and development followed by actual speech files for the recognition experiments).

In 1968 de Boer and Kuyper developed a method known as 'reverse correlation' whereby the impulse response of an auditory nerve can be calculated. If AN fibre discharges are correlated with the input stimulus (e.g. wideband noise), then (assuming a linear system) knowledge of the stimulus prior to the output spike can lead to a quantitative estimate of the impulse response of the unknown system. So a single fibre's revcor function can be measured by averaging the segments of a wideband noise stimulus that precede discharges in the response of the fibre. The Fourier transform of this response is an estimation of the fibre's linear transfer function, and is called the revcor filter. Figure 3.1 shows a schematic diagram of a single section of the auditory nerve model based on the gammatone filter.

![Gammatone Filter Single Section (model 1)](image)

**Figure 3.1** Gammatone Filter Single Section (model 1)

This filter characterises many of the response features of a fibre, including latency and phase and the associated tuning characteristics, for stimuli as complex as a speech input from a single sound source (de Boer and de Jongh, 1968; Carney and Yin, 1988). As expected, the bandwidth of auditory nerve revcor filters increases as a function of input SPL.
This compressive nonlinearity is incorporated into previous models for linear revcor filters (Johannesma, 1972; de Boer, 1975; de Boer and Kruidenier, 1990), by including a feedback loop which continuously varies the filter bandwidth. This is achieved as follows. The gammatone function is an expression which provides a good fit to observed nerve data:

\[
g(t) = \begin{cases} 
\left[\frac{(t - \alpha)}{\tau}\right]^{r-1} e^{-\frac{(t-a)}{\tau}} \cos\left(\omega_{cf}(t - \alpha)\right) & t \geq \alpha \\
0 & t < \alpha 
\end{cases}
\]  

Setting \(\alpha\), (the delay in the onset of the gammatone) to zero, this becomes

\[
g(t) = \left[\frac{t}{\tau}\right]^{r-1} e^{-\frac{t}{\tau}} \cos\left(\omega_{cf}(t)\right) \quad t \geq 0
\]  

The parameters in this equation, along with all the other parameters used in Carney’s single section model, are explained in table 3.1.

The above equation has a simple frequency domain approximation for the parameter range we are interested in:

\[
G(\omega) = \left\{\frac{1}{1 + j\tau(\omega - \omega_{cf})}\right\}^r e^{-j\omega \alpha}
\]  

The parameter \(\tau\) (the time constant of the envelope decay of the gammatone) has an important effect on the bandwidth of the filter. Patterson (1988) has shown that a gammatone filter of order 4 provides a very close fit to human auditory filter shapes.

Carney (1993) describes how to extend Patterson's implementation to include continuous variation of the parameter \(\tau\) via a feedback mechanism. The method takes place in three steps:

1: Shift input data down in frequency by \(\omega_{cf}\).

\[
q(n) = e^{-j\theta_{\omega_{cf}}} P_x(n)
\]

2: Cascade \(\gamma\) low-pass filters, each of the form

\[
H_{lp}(s) = \frac{1}{1 + s\tau}
\]  

29
3: Shift back up in frequency by $\omega_{CF}$.

\[ P_o(n) = \Re \left[ e^{-j \omega_{CF} n} q_v(n) \right] \quad (3.6) \]

where $\Re$ represents the real part of the complex expression.

A schematic diagram of the gammatone filter and the incorporated feedback mechanism is shown in Figure 3.2.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{gammatone_filter.png}
\caption{The gammatone filter with associated feedback}
\end{figure}

The discrete-time expressions for the low-pass filters are derived using the bilinear transform method of digital filter design. In this method, each occurrence of the complex frequency symbol $s$ is replaced by the expression

\[ \frac{2}{T} \frac{(1 - z^{-1})}{(1 + z^{-1})} \quad \text{where} \; T = \text{sampling period}. \]

So substituting this into the transfer function given above gives

\[ H(z) = \frac{1}{1 + \frac{2}{T(1 + z^{-1})} \frac{1 - z^{-1}}{2 \nu_{\max}} + \frac{T(1 + z^{-1})}{T(1 + z^{-1}) + 2 \pi(1 - z^{-1})}} \quad (3.7) \]
\[ H(z) = \frac{Q_{m+1}(z)}{Q_m(z)} = \frac{T(1+z^{-1})}{(T+2\tau) + z^{-1}(T-2\tau)} \]  

(3.8)

\[ Q_{m+1}(z)[(1+2\tau / T) + (1-2\tau / T)z^{-1}] = Q_m(z)(1+z^{-1}) \]

\[ (1+2\tau / T)q_{m+1}(n) = q_m(n) + q_m(n-1) - q_{m+1}(n-1)(1-2\tau / T) \]

\[ q_{m+1}(n) = A[q_m(n) + q_m(n-1) + Bq_{m+1}(n-1)] \]  

where
\[ A = 1 / (2\tau / T + 1); \quad B = 2\tau / T - 1. \]

The coefficient A is fixed for a given fibre. This determines the bandwidth and sensitivity of the filter at low SPL's. The coefficient B is a function of the time-varying feedback signal F(n).

### 3.3 Feedback Control of the Bandwidth

The diagram in Figure 3.2 illustrates the components of the feedback loop. The function of the feedback signal F(n) is to change the bandwidth of the filter as a function of SPL as follows:

- there is no alteration to the bandwidth at low SPL's, i.e. the filter behaves linearly.
- it has an increasing effect between about 30-40 dB and 90 dB.
- Above about 90 dB it saturates, i.e. the filter behaves linearly again.

This is achieved by using the following hyperbolic tangent function:

\[ V_{\text{feedback}} = V_{\text{max}} \frac{1}{(1 + \tanh P_0)} \times \left\{ \tanh\left[0.707P_f(t) / P_{\text{ref}} - P_0\right] + \tanh P_0 \right\} \]  

(3.10)
where the parameter $V_{\text{max}}$ determines the saturation voltage in the depolarising direction, $P_0$ determines the asymmetry of the nonlinearity, and $P_{\text{Dfb}}$ sets the operating point of the nonlinearity. These constants are given empirical values which produce results consistent with those observed for low CF fibres (see section 3.5). The resulting graphs are shown in Figure 3.3.

![Figure 3.3](image1)

**Figure 3.3**  Hyperbolic tangent input-output function and feedback signal

The effects of the feedback signal are not immediately apparent in a change in the bandwidth of the membrane. A suitable time-delay is introduced by including a low-pass filter in the feedback loop. This LPF explains the necessity for an asymmetry in the preceding nonlinearity, since the asymmetry produces a DC component. This means that high frequency components are not attenuated by the LPF. The LPF is implemented as an IIR filter as shown in Figure 3.4.

![Figure 3.4](image2)

**Figure 3.4**  Low pass filter in feedback loop of gammatone filter
\[ V_{\text{out}}(n) = C_1 V_{\text{out}}(n-1) + C_2 [V_{\text{in}}(n) + V_{\text{in}}(n-1)] \]  

(3.11)

where

\[ C_1 = \frac{(C - 2\pi F_c)}{(C + 2\pi F_c)} \]
\[ C_2 = \frac{2\pi F_c}{(2\pi F_c + C)} \]
\[ C = 2 / T \]

The resulting signal \( V_{\text{fb}}(t) \) is scaled and biased to produce the feedback signal \( F(t) \):

\[ F(t) = \frac{3}{2} \tau_0 - \left( \frac{V_{\text{fb}}(t)}{V_{\text{max}}} \right) \left( \frac{\tau_0}{2} \right) \]  

(3.12)

\( \tau_0 \) is a critical value representing the time constant of the exponential damping of the revcor function at a given level. The value for \( \tau_0 \) is obtained from

\[ \tau_0 = C_0 e^{-x(\omega_{CF})/S_0} + C_1 e^{-x(\omega_{CF})/S_1} \]  

(3.13)

where \( C_0, C_1, S_0, S_1 \) are based on empirical observations and \( x(\omega_{CF}) \) is the distance from the apex of the basilar membrane given by Liberman's frequency map (Liberman 1982):

\[ x(\omega_{CF}) = 11.9 \log_{10} \left[ 0.8 + \omega_{CF} / (2\pi 456) \right] \]  

(3.14)

The graph of \( F(t) \) versus input pressure is shown in Figure 3.3.

The operation of the narrow band filter is basically as follows: at low input signal levels the filter behaves like a linear gamma-tone filter. Between about 30 dB and 90 dB the filter output is compressed. Then at higher levels, the filter again behaves linearly.
3.4 Travelling Wave Delay

The output of the filter must be delayed with respect to the input signal by an amount which is dependent on the characteristic frequency of the fibre. The governing equation is

\[ P_{bm}(t) = P_f(t - A_D e^{-\pi(\omega_{CF})/A_L} - 2\pi / \omega_{CF}) \]  (3.15)

Again \( A_D \) and \( A_L \) are empirical constants.

This operation is carried out in software by buffering a number of previous samples, then setting the current sample to the \( n \)-delayth sample, where the value of \( delay \) is given by

\[ delay = \text{(int)}(F_i \times (A_D e^{-\pi(\omega_{CF})/A_L} + 2\pi / \omega_{CF})) \]  (3.16)

3.5 Table of values for Carney's single section model

Table 3.1 lists the various parameters used in the gammatone filter model, including the inner hair cell and synapse model.

The most significant parameters are the input stimulus waveform \( P_x \), the output of the gammatone filter \( P_b \) which is input to the inner hair cell model; and the resulting output voltage which is in turn input to the model of the synaptic connection. It is these latter parts of the model which are now examined.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description (units)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau )</td>
<td>time constant of envelope decay of gammatone</td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td>delay of onset of gammatone (s)</td>
<td></td>
</tr>
<tr>
<td>( \gamma )</td>
<td>order of gammatone</td>
<td></td>
</tr>
<tr>
<td>( \phi )</td>
<td>radian frequency (rad/s) corresponding to CF</td>
<td></td>
</tr>
<tr>
<td>( V_{fh} )</td>
<td>output of feedback nonlinearity (V)</td>
<td></td>
</tr>
<tr>
<td>( V_{max} )</td>
<td>maximum depolarizing hair cell voltage</td>
<td>0.01</td>
</tr>
<tr>
<td>( P_0 )</td>
<td>set 3:1 asymmetrical bias for nonlinearities (rad)</td>
<td>0.462</td>
</tr>
<tr>
<td>( P_f )</td>
<td>output of narrowband filter (Pa)</td>
<td></td>
</tr>
<tr>
<td>( PD_{fb} )</td>
<td>sets operating point of feedback nonlinearity (Pa)</td>
<td>0.005</td>
</tr>
<tr>
<td>( F )</td>
<td>feedback signal (s)</td>
<td></td>
</tr>
<tr>
<td>( \tau_0 )</td>
<td>gammatone time constant estimated at 75 dB for given CF</td>
<td></td>
</tr>
<tr>
<td>( V_{fb} )</td>
<td>output of feedback low pass filter (V)</td>
<td></td>
</tr>
<tr>
<td>( P_x )</td>
<td>stimulus waveform (Pa)</td>
<td></td>
</tr>
<tr>
<td>( x )</td>
<td>distance from apex of basilar membrane (mm)</td>
<td></td>
</tr>
</tbody>
</table>

**Traveling wave delay**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description (units)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{bm} )</td>
<td>narrowband filter output after traveling wave delay (Pa)</td>
<td></td>
</tr>
<tr>
<td>( A_d )</td>
<td>coefficient for traveling wave delay (ms)</td>
<td>8.13</td>
</tr>
<tr>
<td>( A_l )</td>
<td>length constant for traveling wave delay (mm)</td>
<td>6.49</td>
</tr>
</tbody>
</table>

**Inner hair cell**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description (units)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_{in} )</td>
<td>output of IHC nonlinearity</td>
<td></td>
</tr>
<tr>
<td>( P_{dihc} )</td>
<td>sets operating point of IHC nonlinearity</td>
<td>0.001</td>
</tr>
<tr>
<td>( F_c )</td>
<td>cutoff frequency for all IHC lowpass filters</td>
<td>1100</td>
</tr>
<tr>
<td>( V_{ihc} )</td>
<td>intracellular voltage of IHC (V)</td>
<td></td>
</tr>
</tbody>
</table>

**Synapse**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description (units)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_0 )</td>
<td>resting release from synapse (spikes/s)</td>
<td>70</td>
</tr>
<tr>
<td>( S )</td>
<td>output of synapse (spikes/s)</td>
<td></td>
</tr>
<tr>
<td>( C_g )</td>
<td>global concentration (spikes/&quot;volume&quot;)</td>
<td></td>
</tr>
<tr>
<td>( C_l )</td>
<td>local concentration (spikes/&quot;volume&quot;)</td>
<td></td>
</tr>
<tr>
<td>( C_i )</td>
<td>immediate concentration (spikes/&quot;volume&quot;)</td>
<td></td>
</tr>
<tr>
<td>( P_g )</td>
<td>global permeability (&quot;volume&quot;/s)</td>
<td></td>
</tr>
<tr>
<td>( P_l )</td>
<td>local permeability (&quot;volume&quot;/s)</td>
<td></td>
</tr>
<tr>
<td>( P_i )</td>
<td>immediate permeability (&quot;volume&quot;/s)</td>
<td></td>
</tr>
<tr>
<td>( v_l )</td>
<td>local &quot;volume&quot;</td>
<td></td>
</tr>
<tr>
<td>( v_i )</td>
<td>immediate &quot;volume&quot;</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1  Parameters used in Carney's gammatone filter (model 1)
3.6 Hair cell and synapse model

As is discussed in more detail in chapter 5, the pressure from the basilar membrane $P_{bm}(t)$ modulates the current flowing through the hair cell by altering its impedance. The inner hair cell can be modeled as a saturating nonlinearity followed by two low pass filters, a nonlinearity which takes the same form as the hyperbolic tangent function in the feedback loop:

$$V_{out}(t) = \left[ V_{max} / (1 + \tanh P_0) \right] \times \left\{ \tanh \left[ 0.707 \frac{P_{bm}(t)}{P_{Dihe} - P_0} \right] + \tanh P_0 \right\}$$  \hspace{1cm} (3.17)

The two low pass filters are implemented in the same way as that in the feedback loop, i.e. according to Figure 3.4 which gives the difference equation

$$V_{out}(n) = C_1 V_{out}(n-1) + C_2 [V_{in}(n) + V_{in}(n-1)]$$  \hspace{1cm} (3.18)

where

$$C_1 = \frac{(C - 2\pi F_e)}{(2\pi F_e + C)}$$

$$C_2 = \frac{2\pi F_e}{2\pi F_e + C}$$

$$C = \frac{2}{T}$$

The output of the inner hair cell is a voltage which is now used as input to a model of the auditory nerve synapse first proposed by Westerman and Smith in 1988. It consists of two reservoirs and a global source of a neurotransmitter substance, or synaptic material. These reservoirs are connected in series and release material into the synaptic cleft by a diffusion path. The rate of diffusion (i.e. the permeability) is dependent on the input level. The overall output of the model is given by

$$S(n) = P_1(n)C_1(n) \quad \text{spikes/sec.}$$  \hspace{1cm} (3.19)

where

$$P_1(n) = \frac{P_{max} - P_{min}}{V_{Dh} (n) / V_{max} + P_{min}}$$  \hspace{1cm} (3.20)

and

$$C_1(n) = C_1(n-1) + \left[ T / V_1(n-1) \right] \left[ -P_1(n-1)C_1(n-1) + P_2(n-1)\left[ C_2(n-1) - C_1(n-1) \right] \right]$$

$$C_2(n) = C_2(n-1) + \left[ T / V_2(n-1) \right] \left[ -P_2(n-1)\left[ C_2(n-1) - C_1(n-1) \right] + P_C(n-1)\left[ C_C - C_2(n-1) \right] \right]$$  \hspace{1cm} (3.21)

(These equations and their associated parameters are explained in detail in chapter 5).
Thus the overall model accepts as input an arbitrary input and outputs the resulting synaptic discharges for a given fibre. The next step in developing a working model is to expand the single section model into a representation of the overall cochlea.

### 3.7 Adaptation to filterbank formation

Before the overall response of the model to an arbitrary sound source can be modeled it is necessary to arrange a ‘bank’ of individual filters each of which will represent a discrete section of the cochlear membrane. 128 filters are chosen, in order to keep the model consistent with the alternative active models developed by the author in this thesis, which in turn take their inspiration from Ambikairajah’s (1989) 128 channel transmission line model.

![Diagram](image)

**Figure 3.5** Overall synaptic output of the filterbank arrangement of gammatone filters and their associated hair cell/synapses (model 1).
3.8 Tuning characteristics of the filterbank at sample CF

This section presents both the temporal and frequency domain performance of firstly, a single section of the membrane, and then the frequency response of the overall cochlea. In order to get a clear picture of the performance of the various components of the model, a pure sinusoid is first input to a single fibre model. Figure 3.6 shows the temporal response of the model to a sinusoidal input of frequency 400 Hz at 40 dB SPL.

![Diagram of temporal response](image)

*Figure 3.6 Outputs of various stages of the gammatone filter and synapse*
The sinusoidal input is ramped to simulate the effect of the onset of sound. Note the DC bias of the inner hair cell output voltage, which results from equation 3.17. This is necessary because the lowpass filters of the hair cell model have cutoff frequencies below the maximum frequencies of interest, and without this built in DC bias these signals would be attenuated. Note also the units of the feedback signal \( F(t) \). This is controlling \( \tau \), the time constant of envelope decay of the gammatone. Note also the effect of the traveling wave delay. This will increase as the frequency decreases (i.e. as the distance that the traveling wave has to travel along the basilar membrane increases.) Finally note the nature of the synaptic output. It consists of an initial burst of activity which rapidly decreases to a resting or sustained level of firing (see chapter 5).

Figure 3.7 shows the synaptic firing rate for a model fibre with center frequency of 1 kHz. For both responses it can be seen that the model behaves linearly at low SPL's, increases from about 20 dB to 60 dB and then levels off again at higher SPL's. Thus the synaptic model has a dynamic range of about 40 dB which it uses to encode the full input sound level range.

![Figure 3.7](image)

**Figure 3.7** Response of a fibre model to varying input SPL's
This is the expected result due to the feedback nonlinearity and the inner hair cell. Note also that the effect is more pronounced for the sustained rate response than for the onset rate.

Next the tuning characteristics of an individual filter section must be examined. An area rate curve is a plot of the synaptic firing rate versus frequency for a fibre tuned to a particular CF. By varying the SPL's at which a particular fibre responds to the full range of frequencies the tuning characteristics of the fibre as a function of input SPL can be observed.

Figure 3.8 shows the area rate curves for a filter section tuned to 1 kHz. Note the low response level near the threshold of hearing, increasing to a reasonably sharply tuned bandwidth at intermediate frequencies. At higher input levels the area rate curves flatten out somewhat, reflecting the damping effect of the feedback nonlinearity in the gammatone filter section.

![Area rate curves for fibre with CF 1 kHz](image)
Finally, Figure 3.9 shows the output from the entire cochlear model consisting of 128 individual sections. The center frequencies are distributed according to the Mel frequency scale. The sinusoidal input is presented simultaneously to all filter sections. The fibres fire rapidly at the channel number with center frequency corresponding to the applied sinusoid. Figure 3.10 shows the corresponding diagram when five tones are applied simultaneously.

![Graphs](https://via.placeholder.com/150)

**Figure 3.9**  Filterbank response to an individual sinusoid

The six peaks of synaptic activity are clearly visible when the combination of tones is applied to the model.

![Graph](https://via.placeholder.com/150)

**Figure 3.10**  Six simultaneous tones applied to filterbank
3.9 Summary

This chapter has presented a detailed overview of a single channel model of the auditory periphery, based on the work of Carney (1993). The model's most important feature is the active nonlinear feedback mechanism simulating the properties of the outer hair cells of the cochlea, which amends the otherwise linear estimate of the behaviour of the middle ear and basilar membrane provided by the reverse correlation function. This model is amended to work as a parallel filterbank, and will be used as a benchmark against which to measure the performance of the alternative compressive models which are developed in this work.
Chapter 4
Development of alternative active models (Models 2 and 3)

4.1 Introduction

This chapter describes the development of two alternative, active, nonlinear models of the basilar membrane, in relation to that described in chapter 3. This is desirable because the gammatone filter is relatively inefficient from a signal processing point of view compared with other commonly used membrane filter models. If real time implementations of the speech processing models developed in this thesis are to be found it is essential that the front end processor be as efficient as possible. This means simplifying the model as much as possible while retaining it's salient features. The inner hair cell and synapse models are the same as those used before, in order to compare the two membrane models. In the development of the new models particular attention is paid to the following behavioural characteristics:

- Temporal response.
  The new model must produce similar fine-time outputs to the gammatone filter. As before, these consist of a burst of activity at a tone onset which gradually settles down to a resting rate for a tone at a particular decibel level.

- Level dependence.
  The bandwidth of the membrane filter must increase at high input levels. Also, there should be a gain for low input signals, which decreases as the signal level is increased.

- Neural tuning curves.
  The output of a particular channel should peak at it's centre frequency - i.e. the frequency it's tuned to. Particular emphasis is to be placed on improving the selectivity of this model over the previous one.
It will be seen that two models are eventually arrived at. Firstly a single channel alternative to Carney's model is developed which has similar characteristics according to the three criteria listed above (i.e. with regard to its compressive capabilities and level dependent bandwidth), while being considerably less computationally intensive. It consists of two filters, a constant Q bandpass filter followed by a variable Q second order lowpass filter.

However it will be seen that single channel models such as these two have an inbuilt problem with regard to the amount of compression they are able to perform.

Using a combination of the techniques developed for the above single channel model, combined with the transmission line model of Ambikairajah (1989) much more realistic compression ratios can be attained. By accumulating the gains of individual filter sections over the cascade of sections compression ratios of almost 2.5:1 can be obtained without over-compromising either latencies or bandwidth.
4.2 Model 2: Alternative parallel filterbank

4.2.1 Basic plan for alternative single channel model

The principal idea for the new model is shown in Figure 4.1. The action of the basilar membrane is to be modelled by a bank of narrow band filters simulating the frequency selectivity of the membrane. In order to control the bandwidth of the membrane at a particular point, a feedback mechanism is proposed to control the Q (centre frequency / bandwidth) of the filter. Since the centre frequency remains constant, varying the Q will change the bandwidth of the filter.

*Note:* It is the output of an outer hair cell model which is averaged rather than the filter itself. This is left out of the above diagram for simplicity. The actual feedback loop is as shown in Figure 4.2.

Before the feedback loop is connected, the maximum and minimum energies of the filter outputs must be calculated. The energy is taken as the sum of the outputs squared over the integrating period - in this case 1 ms. The minimum energy corresponds to the filter output at 0 dB (taken at 95 µPa), and the maximum energy was taken at 100 dB.
4.2.2 Digital Filter Design

The transfer function of the chosen filter is:

\[ H(s) = \frac{b_p s}{s^2 + b_p s + \omega_p^2} \]  (4.1)

The digital equivalent is obtained by using the bilinear theorem, which maps the analogue domain to the digital according to the equation

\[ s \rightarrow \frac{2(1-z^{-1})}{T(1+z^{-1})}, \]  (4.2)

giving

\[ H(z) = \frac{b_p 2 (1-z^{-1})}{T (1+z^{-1})}, \]  (4.3)

\[ H(z) = \frac{b_p 2 (1-z^{-1})}{4(1-2z^{-1} + z^{-2}) + (2b_p - 2b_p z^{-1})T(1+z^{-1}) + \omega_p^2 T^2 (1+z^{-1})^2}, \]  (4.4)

\[ H(z) = \frac{2b_p - 2b_p T z^{-2}}{(4+2b_p T + \omega_p^2 T^2) + (2\omega_p^2 T^2 - 8)z^{-1} + (4-2b_p T + \omega_p^2 T^2)z^{-2}}, \]  (4.5)

\( b_p \), \( \omega_p \), and \( T \) are parameters that need to be chosen based on the requirements of the system.
and finally

\[ H(z) = \frac{a_0 - a_z}{1 + b_1 z^{-1} + b_2 z^{-2}}. \]  

(4.6)

where

\[ a_0 = a_1 = \frac{2b_p T}{4 + 2b_p T + \omega_p^2 T^2}, \]  

(4.7)

and

\[ b_1 = \frac{2\omega_p^2 T^2 - 8}{4 + 2b_p T + \omega_p^2 T^2}, \quad b_2 = \frac{4 - 2b_p T + \omega_p^2 T^2}{4 + 2b_p T + \omega_p^2 T^2}. \]  

(4.8)

For a given filter with centre frequency \( CF_p \), the coefficients can be set if the value of \( Q \) is known, from the relationship between \( b_p \) and \( Q \) : \( b_p = \omega_f / Q \). The value of \( Q \) is first set to its maximum value, and the coefficients are calculated. Then, every millisecond the value of \( Q \) is changed in response to the signal level and new coefficients calculated. It should be observed that the bandwidth of an individual filter will broaden considerably at higher signal input levels.

The first test to be performed on the model checks the \( Q \) control capabilities of the feedback loop. In all cases (unless otherwise stated) the tests are carried out on the filter tuned to 1880 Hz (filter no 40). Figure 4.3 shows the variation of \( Q \) in response to an input signal which increases in 4 equal steps and then linearly from 0 to 100 dB. The filter \( Q \) is shown to rapidly follow the input signal level as it changes.

Next the effect which the \( Q \) has on the latency and bandwidth of the filter is shown (Figure 4.4). At low \( Q \) the impulse response of the filter lasts for about 1 ms and the bandwidth of the filter is approximately 1880 Hz. These figures change to 4 ms and 500 Hz respectively for the high \( Q \) case. (These results are obtained by disconnecting the feedback loop and fixing the \( Q \) to its minimum and maximum values respectively).
Figure 4.3  Change in $Q$ with input SPL

Figure 4.4  Impulse response and frequency response of a filter at low and high $Q$ values.
These results show that the principal of controlling the filter bandwidth as a function of the input signal level can be achieved relatively straightforwardly. The feedback nonlinearity, as in the gammatone model, is simulating the action of the outer hair cells of the cochlea in broadening the bandwidth of the basilar membrane at high SPL's.

**4.2.3 Inclusion of compressive mechanism into the model**

The next characteristic to be incorporated into this new design is the compression of the input signal into a smaller range. The ear must cope with an enormous variation in pressure from 0 dB SPL up to the threshold of pain, a variation which spans many orders of magnitude. It is unfeasible to suppose that the auditory nerves can directly code this variation in signal intensity into neural discharges.

Evidently some degree of signal compression is required. The gammatone model provides about 20 dB of compression over a 100 dB input range. More recent models have achieved higher compression ratios (e.g. Kates 1993). The aim here is to try to match, and if possible improve on the compression inherent in the gammatone model, i.e. an approximately 1.66:1 compression ratio. This means that the filter model must be capable of providing a level dependent gain factor of 40 dB at low input levels, dropping to unity gain at higher levels (100 dB).

![Output signal (dB) vs Input signal (dB)](image)

**Figure 4.5 Desired gain characteristics of the cochlear filter**

The first thing which must be noticed is that the band pass filter used to model the variable bandwidth membrane has no inherent gain control. This follows from the equation:
Replacing $s$ with $j\omega$ gives

$$H(\omega) = \frac{j\omega \frac{\omega_p}{Q}}{-\omega^2 + j\omega \frac{\omega_p}{Q} + \omega_p^2}$$  \hspace{1cm} (4.10)$$

$$|H(\omega)| = \frac{\omega \frac{\omega_p}{Q}}{\sqrt{(\omega_p - \omega)^2 + \left(\frac{\omega \omega_p}{Q}\right)^2}}$$  \hspace{1cm} (4.11)

And at $\omega=\omega_p$

$$|H(\omega)| = 1$$

Thus varying $Q$ will not vary the gain of the filter at centre frequency.

The following analysis shows that a simple second order lowpass filter has a gain of $Q$. The analogue equation for the filter is:

$$H(s) = \frac{b_p s}{s^2 + b_p s + \omega_p^2}$$  \hspace{1cm} (4.12)$$

where $\omega_p$ is the resonance frequency and $b_p$ the bandwidth of the filter respectively.

This has the magnitude response

$$|H(\omega)| = \frac{\omega_p^2}{\sqrt{(\omega_p - \omega)^2 + \left(\frac{\omega \omega_p}{Q}\right)^2}}$$  \hspace{1cm} (4.13)$$

And at $\omega=\omega_p$

$$|H(\omega)| = Q$$
If $Q_{\text{max}}$ and $Q_{\text{min}}$ are set to 10.0 and 1.0 respectively, then the filter gain will drop from 20 dB at low input levels to unity gain at high input levels.

Note: The lowpass characteristic of this filter has no ramifications for the eventual frequency selectivity of the model, since the filters will be implemented in a coupled filterbank formation. This is to simulate the mechanical coupling of adjacent sections of the membrane, and will result in a sharpening of the lowpass slopes of the filter shape.

The bilinear transform is used to obtain the corresponding digital filter as follows:

\[
H(s) = \frac{\omega_p^2}{s^2 + b_p s + \omega_p^2} \tag{4.14}
\]

\[
H(z) = \frac{\omega_p^2}{\left[2 \left(1 - z^{-1}\right)^2\right] + \frac{b_p}{T} \left(1 + z^{-1}\right) + \omega_p^2} \tag{4.15}
\]

\[
H(z) = \frac{\omega_p^2 T^2 (1 + z^{-1})^2}{4(1 - z^{-1}) + 2b_p T (1 - z^{-1})(1 + z^{-1}) + \omega_p^2 T^2 (1 + z^{-1})^2} \tag{4.16}
\]

\[
H(z) = \frac{\omega_p^2 T^2 (1 + 2z^{-1} + z^{-2})}{(4 + 2b_p T + \omega_p^2 T^2) + (\omega_p^2 T^2 - 8)z^{-1} + (4 - 2b_p T + \omega_p^2 T^2)z^{-2}} \tag{4.17}
\]

\[
H(z) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}} \tag{4.18}
\]

where

\[
a_0 = a_1 / 2 = a_2 = \frac{\omega_p^2 T^2}{(4 + 2b_p T + \omega_p^2 T^2)} \tag{4.19}
\]

and

\[
b_1 = \frac{(\omega_p^2 T^2 - 8)}{(4 + 2b_p T + \omega_p^2 T^2)}, \quad b_2 = \frac{(4 - 2b_p T + \omega_p^2 T^2)}{(4 + 2b_p T + \omega_p^2 T^2)}. \tag{4.20}
\]
Figure 4.6 shows the effect of varying Q on the frequency response of the filter. It can be seen that the gain of the filter increases along with bandwidth. Now low amplitude signals receive a boost via the filter gain. Note the increasing latencies as filter Q increases, from less than 1ms at minimum Q to about 6ms at maximum Q. The filter shape changes from sharply resonant at high Q to low pass at low Q.

Figure 4.6 Impulse and frequency responses of the 2nd order lowpass filter at varying Q's

Figure 4.7 shows the compression of the output signal with respect to the input signal over the range 0 to 100 dB SPL. Note the 20 dB of compression over the input range, a similar result to that of the gammatone filter. Figure 4.8 shows the area rate curves for the model when the same model of the inner hair cell and synapse that was used in Carney's model is included in the present model. This test was carried out on an isolated filter section. The lack of adjacent
filter coupling is visible in the low pass nature of the resulting plots. The low frequency components will be removed when the full filterbank is running.

In order to achieve a much higher compression ratio consider the arrangement shown in Figure 4.9.

**Figure 4.7** Input/Output

**Figure 4.8** Area rate curves for lpf model

**Figure 4.9** Include direct control of gain factor K in feedback loop
In this model the same signal that controls the filter Q is used to control a gain factor which is placed after the feedback loop. The gain is set to drop exponentially from 60 dB (1000) at low input signals to zero at higher signals, thus providing a compression ratio of 2.5:1. However this arrangement is found to compromise the performance of the model in another crucial respect, that of the frequency selectivity of the model, for the reason apparent from Figure 4.10.

The gain control increases the filter gain in response to signal level irrespective of frequency. Thus the gain of the filter will be slightly higher away from centre frequency than it is at the centre frequency, as shown in Figure 4.10(d). This has the effect of increasing the bandwidth of the filter, which is the opposite effect to that which we want (Figure 4.10(c)). Although this method enables the input signal to be compressed over the entire frequency range, it is not pursued, as the frequency selectivity requirements are excessively compromised.

Figure 4.10  Effect of direct gain factor control on bandwidth
The desired compression must be obtained as an inherent characteristic of the filter, making the gain control frequency dependent. This leads on to the next attempt to increase the gain factor.

The next arrangement considered is to use two 2nd order low pass filters in cascade in each filter channel (Figure 4.11). This should double the maximum gain from 20 to 40 dB at low signal levels. The graphs in Figure 4.12 show the test results on a single filter channel. The 40 dB gain at low SPL levels gives an improved compression ratio of 1.66:1. However, the effects of having two filters, both with Q’s set to their maximum values of 10 leads to unacceptable delays to the impulse response of the filter. The output of the second filter is still perceptably oscillating after 10 ms - much too long for a realistic model of the fine time response of an auditory channel.

![Figure 4.11 Two 2nd order bandpass filters in series](image-url)
Figure 4.12  Response characteristics for two 2nd order lowpass filters in series

4.2.4 Final version of model 2

Figure 4.13  Two 2nd order bandpass filters in series (model 2)
The arrangement in Figure 4.13 is the final version of this model. It gives a reasonable compromise between compression and latency. The signal is first passed through a bandpass filter, tuned to centre frequency. The output of this filter is then used to control the Q of a 2nd order low pass filter, tuned to the same frequency.

**Figure 4.14** Response characteristics for the final single channel model

Figure 4.14 shows the impulse response for this filter arrangement at low and high Q values. The latency of the filter is still acceptable, at less than 10 ms, and the bandwidth of the channel is narrower than all the previous models, even at low Q.
4.2.5 Evaluation of model 2 performance

Figure 4.15 Area rate curves for final model

Figure 4.15 shows the area rate curves of the filter outputs for the channel tuned to 1880 Hz. Figure 4.16 shows the area rate curve with the inner hair cell / synapse model of Carney included. With reference to Figure 3.8 it can be seen that the selectivity of the current model is superior to the gammatone filter at all SPL's, and particularly at higher input levels.

Figure 4.16 Area rate curves

Finally a series of these filters is arranged in a filterbank formation (Figure 4.17), as for the gammatone filter. Again there are 128 channels tuned to frequencies according to the Mod-Log frequency scale.
Note the double spatial differentiation to model the mechanical coupling of the basilar membrane.

Figure 4.17 Filterbank formation with spatial differentiation

Figure 4.18 Frequency response of filterbank
The graph in Figure 4.18 shows the frequency responses of every 4th filter in the filterbank. The scaling factors were obtained by calculating the frequency responses of all the filters, and normalising the magnitudes of all of them by inverting the peak value of each response. When the scaling factors are included the peak of each response is the same.

To test the frequency selectivity of the model 5 sinewaves are input to the model, and the filter outputs every millisecond are superimposed. Figure 4.19(a) shows the filter outputs for a signal input at 0 dB. Note the sharpening effect which spatial differentiation has on the tuning characteristics of the model. Figure 4.19(b) shows the same plots at 100 dB. As expected the bandwidth increases substantially at the higher level.

Figure 4.19  Filterbank output in response to 5 tones

Figure 4.20 shows the waveform of a speech sample. When this waveform is normalised to two different levels (40 dB and 80 dB) and input to the system, the synapse outputs
(superimposed every 2 ms) are as shown in Figure 4.21. The fine time responses of 4 channels are as given in Figure 4.22.

Figure 4.20  Sample speech pattern

Figure 4.21  Superimposed filter outputs in response to speech sample
Figure 4.22  Fine time filter outputs to spoken 'p'
4.3 Transmission line adaptation for realistic compression ratios (model 3)

4.3.1 Proposal for realistic compression ratios in a cochlear model

It is apparent from the models and governing equations considered in the previous section that single section models are insufficient to encapsulate the dynamics of the three performance criteria that are of interest to this study. In an attempt to dramatically increase the overall cochlear gain an adaptation of the basic transmission line model of Ambikairajah (1989) is proposed along the lines used by Kates (1993) (see chapter 2). It will be seen that this method results in a realistic compression ratio while maintaining bandwidth and fine time performance.

The filters are arranged in cascade, with the output of each filter being fed to a second filter which is tuned to the same frequency. The outputs of the second filters are sent to a peak detection mechanism which controls the Q values of the appropriate filters. It will be seen that this arrangement enables an almost uniform 60 dB of compression over most of the cochlear filters to be obtained. This high compression ratio is possible because the signal at any given tap has been boosted by all the preceding filter gains. The low Q values used also mean that realistic latencies are achieved (between 10 and 20 ms for the 1 kHz channel). Figure 4.23 shows a single section of the basilar membrane model and its associated transfer functions.

\[
\begin{align*}
\frac{v_p(s)}{v_1(s)} &= \frac{\omega_p^2}{s^2 + B_p s + \omega_p^2} \\
Q_p &= \frac{\omega_p}{B_p} \\
Q_p &= \frac{\omega_p}{B_p}, \\
Q_z &= \frac{\omega_z}{B_z}
\end{align*}
\]

![Figure 4.23](image.png)

**Figure 4.23** Single section of the digital transmission line model of the basilar membrane
4.3.2 Modified Transmission Line (MTL) Model

In the modification proposed by the author, a second filter has been added at each tap to enhance the gain and selectivity of the model. The fact that the filters used must have a variable gain (that is the gain must change with bandwidth) places a constraint on the choice of filter as discovered in the development of a single section model. As in that case a resonant second order lowpass filter is used, in which the Q value directly controls the gain.

\[ H(s) = \frac{\omega_p^2}{s^2 + B_p s + \omega_p^2}, \quad \Rightarrow |H(\omega)|_{\omega=\omega_p} = Q \] (4.21)

Thus, if we declare all of the filter Q's to have a minimum value of 1 (this is the passive state), then the entire cochlea will be zero gain. This also eliminates the need for the constant multipliers \( K, G_0, G_p \) and \( G_z \) of Figure 2.11. The digital filters are designed using the bilinear transform. At each tap the displacement is equalled to the pressure value, and this is used as an input to the second filter. The pressure transfer function is given below:

\[ \frac{V_o(s)}{V_i(s)} = \frac{\omega_p^2}{s^2 + B_p s + \omega_p^2} \frac{s^2 + B_z s + \omega_z^2}{\omega_z^2} \] (4.22)

where \( \omega_p \) and \( \omega_z \) are the pole and zero frequencies and \( B = \omega_z/Q \).

Applying the bilinear transformation the digital transfer function becomes

\[ \frac{V_o(z)}{V_i(z)} = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}} \] (4.23)

with coefficients

\[ a_0 = \frac{\omega_p^2 + b_z K + K^2}{D}, \quad a_1 = \frac{2\omega_z^2 - 2K^2}{D}, \quad a_2 = \frac{\omega_z^2 - b_z K + K^2}{D}, \]
\[
\begin{align*}
\dot{\theta} &= \frac{2\omega \dot{\theta} + 2K}{D}, \quad \ddot{\theta} = \frac{-2\theta - 2\dot{\theta}}{D}
\end{align*}
\]  
(4.24)

where \[K = \frac{2}{F_s}\] and \[D = \omega_p^2 + b_pK + K^2\].

It is important that digital prewarping is used for filters designed using the bilinear transform, to compensate for the nonlinear analog to digital frequency mapping at high frequencies. This is compensated for by prewarping analog frequencies according to

\[
\theta = 2\tan^{-1}\left(\frac{\omega T}{2}\right)
\]
(4.25)

The resulting structure of a single filter section is simpler to that of Figure 4.23 and becomes:

![Diagram of single section of modified transmission line model]

**Figure 4.24** Single section of the modified transmission line model

When the second filter and Q control structures are included, the overall cochlear model looks like that shown in Figure 4.25. Before the pressure outputs from the first filters are input to the second, there is a spatial differentiation stage. This represents the mechanical coupling of adjacent sections of the basilar membrane, and serves to sharpen the response of each section. A simple subtraction serves as a differentiator:

\[
D(i) = D(i) - D(i-1) \quad \text{for } i = 1..128
\]

A scaling factor is included at this point to ensure that when all the filters are locked in the low Q (passive) condition the frequency responses of each channel has a 0 dB peak (Figure 4.27). The scaling factor is set by applying an impulse response at the input to the model and taking
an FFT of the output at each channel. Inverting the peak magnitude of each response yields a set of scaling factors.

![Diagram of Modified Transmission Line Model of the Cochlea (Model 3)](image)

**Figure 4.25** Modified Transmission Line Model of the Cochlea (Model 3)

The second filter used at each tap to enhance the gain and selectivity of the model is of the same form and is tuned to the same frequency as the corresponding filter in the cascade structure (Figure 4.24). The second filter output is to be used to control the filter bandwidths and gain. That is, at low signal levels the gain is high and the bandwidth is low, and at high signal levels the gain is low and the bandwidth is broader. This dynamic in the gain and bandwidth of the filters is obtained by adjusting the Q's of each filter (Kates[6]) as a function of the peak output from the second filters. The maximum Q values ($Q_{pl}$) of the cascaded filters vary from 1.2 to 8.0 from the low to high frequency ends, on a linear cochlear distance scale. The relationship between the zero and pole parameters are: $\omega_p(i) = 1.15 \omega_p(i-2)$ ($i = \text{filter}$)
number, $\omega_{z1} = \text{zero resonant frequency}$, $\omega_{p1} = \text{pole resonant frequency}$), and $Q_m = 1.50Q_n$. For the second filter the maximum Q values ($Q_{p2}$) are also frequency dependent, determined by the relationship $Q_{p2} = 1.5(1+f)$ with $f$ in kHz, and the zero/pole parameters vary according to $\omega_{z2} = \omega_{p2}/2$ and $Q_{z2} = 2.0Q_{p2}$.

The feedback control mechanism works as follows: the outputs of each of the second filters are sent to a peak detection mechanism. This detects the peak output over a span of 8 filters either side of the one whose Q is to be controlled. This peak output is then used to determine the ratio of peak output to maximum output ('$r$'), which in turn controls the Q setting of each filter. The cascaded filter Q’s are given by $Q_{p1} = (0.3 + 0.7r) \times Q_{1\text{max}}$ and the second filter Q’s are given by $Q_{p2} = (0.1 + 0.9r) \times Q_{2\text{max}}$. Filter outputs should lie between 40 and 80 dB, to ensure that the 40 dB range of filter outputs matches the 40 dB dynamic range of the inner hair cell/synapse model which is used in this work. This is why the frequency responses are scaled to peak at -20 dB at low Q.

![Magitude response (Arbitrary scale)](image)

**Figure. 4.26** Gain accumulation in the cascaded model

Figure 4.26 shows diagramatically why a cascaded model of the cochlea is essential for a large amount of compression. It is because a particular filters’ gain is dependent on the frequency responses of all the preceding filters. An individual filter by itself can have up to 20dB of variable gain (say for example that the Q of one of the filters in this model were varying between 1 and 10). But using a transmission line model each individual filter can vary over a
much smaller range while giving a much more dramatic overall effect. It is desirable to minimise the variation in the individual filters in order to keep variations in stability and channel latencies down.

4.3.3 Evaluation of performance of MTL

Figure 4.27 shows the frequency responses of every 6th filter when all the Q’s (Q$_{q1}$ and Q$_{q2}$) are fixed in both the low and high states (Figure 4.27(a) and (b) respectively). In the high Q state, both the low and high frequency responses do not have a full 60 dB gain. This is due to the fact that for the high frequency filters there are not enough preceding filters for the gain to have accumulated to 60 dB, while for the low frequency case it is because the overall gain contributions of the preceding filters at these frequencies does not sum to 60 dB. Finally Figure 4.28 shows the superimposed outputs for 5 tones applied at about 30 dB SPL. At this fairly low input level the Q values are quite high and thus there is good frequency resolution evident in the model.

![Frequency response of every 6th filter](image)
Figure 4.28  Superimposed outputs every ms for 5 sinewaves applied at 30 dB SPL

4.4 Summary

This chapter has given a detailed account of the steps taken to develop a model of the auditory periphery which produces realistic behaviour with regard to the active compression which exists in the cochlea, while maintaining the crucial characteristics of finely tuned cochlear sections and realistic latencies. It is shown that it is not feasible to produce a single channel model of the cochlea which results in substantial compression of the input range of signals. Realistic compression can only be achieved by using the accumulative gain of a cascade structure of filters.

The two models produced here can now be compared with the earlier described gammatone model, in terms of performance under test conditions, and then as a front end-processor to a neural network based speech recognition system.
Chapter 5
Inner Hair Cell/Synapse model

5.1 Introduction

In the previous chapters the development of models of the basilar membrane is explained. The membrane filter sections result in 128 element vector outputs which represent the displacement or velocity of the mechanical motion of the basilar membrane. This mechanical motion is modulated into an electrical current in the cilia or the inner hair cells which line the basilar membrane (by means of the organ of Corti). This can be represented by a straightforward lowpass filter.

The next stage in the processing of the input pressure wave into information which can be transmitted along neurons to the brain is a further transduction stage in which the electrical energy in the IHC's is converted into auditory-nerve fiber discharges, or spikes. This mechanism takes place at the synapse between the IHC's and the auditory nerve, and is effected by the electrical signal from the inner hair cell causing the release of a neurotransmitter substance into the synaptic cleft between the two structures. The neurotransmitter causes the generation of a pulse train in the efferent nerve fiber. These discharges are the means by which information about a stimulus is transmitted to the higher centres of auditory processing.

The behaviour of auditory nerve fibers in response to stimuli has been found to mirror that of that part of the basilar membrane to which they are responsive, i.e. to which they connect by means of the inner hair cells. Thus, the tuning curves for single fibers display a similarity to the mechanical tuning curves of that section of the cochlea, in that there seem to be two mechanisms contributing to the fibre tuning curve, a broadly tuned mechanism, and a more sharply tuned mechanism which contributes a sharp tip to the tuning curve, and which is
susceptible to physiological damage. These two tuning mechanisms are assumed to be passive and active respectively. It has also been noted that there is a wide variation in the tuning characteristics of individual fibres (Evans 1975), for example those with high CF's tend to be more sharply tuned than those in the apical (towards the end) part of the cochlea. Also the shape of the tuning curve varies along the length of the basilar membrane.

The detailed time course of the response of auditory nerve fibers has been measured by many researchers (e.g. Kiang et al. 1965), and a number of techniques have been developed to represent it. A good example is the Post Stimulus Time (PST) histogram, which involves splitting the time course of the response into a number of short time intervals, or bins, and incrementing a histogram value for each bin in which a discharge occurs. This is carried out for a large number of stimulus presentations to reduce the statistical variation of the measurements.

The "spontaneous rate" is the rate at which most of the auditory fibers discharge at random in the absence of an input stimulation, and it can vary from just above 0 spikes/sec to over 100 spikes/sec. The encoding of acoustic stimuli by the nervous system must require the participation of a large number of nerve fibers, since the rate at which single auditory nerve fibers can discharge is limited to a few hundred spikes per second, too low for any individual fiber to encode such a waveform (Geisler, 1988).

"Adaptation" is the name given to the phenomenon whereby the discharge rate of an auditory nerve fiber settles down, after an initial period of high activity, to a steady resting rate after a period of time in response to a stimulus with a sudden onset. Rapid adaptation refers to the period immediately (i.e. a few milliseconds) after the onset of the stimulus in which the rate of decay is highest. Short term adaptation follows, in which the rate of decay is more gradual and extends over several tens of milliseconds. Finally, after removal of the stimulus, the firing rate drops below the spontaneous rate, to gradually recover to the spontaneous rate. Figure 5.1 shows a typical auditory fiber response to a tone burst.
Models that generate detailed and realistic temporal responses (i.e. discharge rates) in response to stimuli are a necessary prerequisite to detailed modeling of the processes which occur at higher levels of the auditory system, e.g. the cochlear nucleus, and the incorporation of such models into speech processing systems. Because speech, by its nature, is a non-stationary dynamic process, temporal evolution of the signal may be just as important to the higher centres of speech processing as the spectral behaviour, thus the importance of reliable and accurate models of the neural response. The development of models of the neural adaptation process are dependent on the physiological experiments of, amongst others, Schroeder and Hall (1974), Cooke (1986), Meddis (1986,1988).

This chapter gives a description for comparative purposes of several proposed models for transduction, followed by a detailed description of the neural transduction model of Westerman and Smith (1988) used with the various models presented in this thesis.
5.2 Transduction models

*Seneff's Adaptation Model*

This model was presented by Seneff in 1988 as part of an overall auditory periphery model, in which the membrane displacement output from a number of critical band filters are processed by a four stage nonlinear transduction mechanism. The four stages are shown in Figure 5.2 and are a saturating half-wave rectifier, a short-term adaptation mechanism, a low-pass filter and a rapid automatic gain control (AGC).

![Diagram of Seneff's Adaptation Model](image)

Each of the four stages has a corresponding physiological phenomenon, as shown in the diagram.

The half wave rectifier is represented by the equation

\[
\begin{align*}
y &= 1 + A \tan^{-1}(Bx) \quad x > 0 \\
y &= e^{dx} \quad x \leq 0
\end{align*}
\]  

(5.1)
The model for short-term adaptation consists of two independent mechanisms which influence the release of neurotransmitter into the synaptic cleft. It flows from a “source”, across a “membrane”, the concentration gradient across which controls the rate of flow. This behaviour is represented by the equation:

\[
\frac{dC(t)}{dt} = \mu_a [S(t) - C(t)] - \mu_b C(t) \quad C(t) < S(t)
\]

\[
= -\mu_b C(t) \quad C(t) \geq S(t)
\]

(5.2)

where \(S(t)\) is the transmitter concentration in the source region, and \(C(t)\) is the concentration in the pre-synaptic cleft. The constants \(\mu_a\) and \(\mu_b\) control the rate of flow of neurotransmitter across the membrane, and flow due to natural decay respectively.

The output of the adaptation stage is passed through a lowpass filter to simulate high-frequency synchrony reduction, and finally the neurotransmitter concentration enters the rapid Automatic Gain Control represented by the equation

\[
y(n) = \frac{x(n)}{1 + K_{AGC} \langle x(n) \rangle_t}
\]

(5.3)

where \(K_{AGC}\) is a constant, and \(\langle x(n) \rangle\) symbolises the expected value of \(x(n)\), obtained by integrating \(x(n)\) over a period \(\tau\).

**Schroeder Hall Adaptation Model**

The Schroeder Hall reservoir model (Schroeder and Hall, 1974) assumes that there is a neurotransmitter source for each basilar membrane filter output which, by means of the concentration of the substance, controls the firing rate. The model also allows for spontaneous firing as well as exponential decay of transmitter.

The rate of change of neurotransmitter is given by

\[
\frac{dn(t)}{dt} = A - [S_o + S_b + Dq(t)]n(t)
\]

(5.4)
This model, as implemented by Cohen (1989) resulted in a very efficient transduction process.

\[ f(t) = \left[ S_0 + Dq(t) \right] n(t) \]  

(5.5)

This model, as implemented by Cohen (1989) resulted in a very efficient transduction process.

**Meddis' Adaptation Model**

In the adaptation model of Meddis (1986, 1988) the transmitter substance \( q(t) \) is actually stored in the hair cell and leaks into the synaptic cleft at a rate which is proportional to membrane permeability, which in turn is controlled by the basilar membrane displacement \( s(t) \). A schematic diagram of Meddis' proposal is shown in Figure 5.3.

**Figure 5.3  Adaptation model of Meddis (1986)**
The transmitter in the cleft, $c(t)$, is mostly returned to the reprocessing store ($w(t)$) in the hair cell at a rate $rc(t)dt$. Transmitter is lost from the cleft at a rate $lc(t)dt$. The store of free transmitter is replenished at a rate $y[M - q(t)]$. The equation describing the rate of change of the hair cell contents is

$$\frac{dq(t)}{dt} = y[M - q(t)] + xw(t) - k(t)q(t)$$  \hspace{1cm} (5.6)$$

where

$$k(t) = \begin{cases} \frac{g[s(t) + A]}{s(t) + A + B} & s(t) + A > 0 \\ 0 & s(t) + A \leq 0 \end{cases}$$  \hspace{1cm} (5.7)$$

$$\frac{dc(t)}{d(t)} = k(t)q(t) - lc(t) - rc(t)$$  \hspace{1cm} (5.8)$$

$$\frac{dw(t)}{dt} = rc(t) - xw(t)$$  \hspace{1cm} (5.9)$$

As is shown by Jones (1994), these models all have approximate dynamic ranges of about 40 dB, similar to that of the model examined in the next section, that proposed by Westerman and Smith (1988). This model is the one eventually adopted, in the absence of any clear advantage of one model over another, in order to provide accurate comparisons with the results of Carney's gammatone filter model (1993),
5.3 The inner hair cell

This section presents the inner hair cell model chosen in conjunction with the synaptic model of the next section, for use in this work. The pressure from the basilar membrane modulates the current flowing through the hair cell by altering its impedance. The simplest type of hair cell model is a simplification of that of Davis (1953), and is shown in Figure 5.4.

![Figure 5.4 Simple model of the inner hair cell](image)

This model represents the bending of the inner hair cell cilia due to the spatially differentiated mechanical displacement of the basilar membrane. The spatial differentiation represents the mechanical coupling between the cochlear fluid and the inner hair cell cilia. Half wave rectification is included since the modulation of current flow is due to the cilia bending in one direction only. The remainder of the model consists of a half wave rectifier followed by a low pass RC filter.

![Figure 5.5 Inner hair cell model with hyperbolic tangent nonlinearity](image)
Alternatively the inner hair cell can be modeled as a saturating nonlinearity followed by two low pass filters, (see Figure 5.5).

In this case input/output behaviour for inner hair cells has the form of an asymmetrical nonlinearity. The two low pass filters represent the electrical filtering of the inner hair cell membrane. The nonlinearity which relates the output voltage of the inner hair cell $V_{in}(t)$ to the output of the cochlear filter $P_{tm}(t)$ takes the form of the hyperbolic tangent:

$$V_{in}(t) = \left[ \frac{V_{max}}{1 + \tanh P_0} \right] \times \left\{ \tanh \left[ 0.707 \frac{P_{tm}(t)}{P_{Dihc}} - P_0 \right] + \tanh P_0 \right\} \tag{5.10}$$

![Figure 5.6 Hyperbolic tan nonlinearity](image)

$V_{max}$ and $P_0$ set the maximum output voltage and the asymmetry of the nonlinearity respectively, and are set according to empirical observations of the behaviour of inner hair cells with regard to response amplitudes and asymmetry. The setting of $V_{max}$ results in maximum DC and AC response components of 3 and 10 mV respectively (Dallos 1985, 1986). The setting of $P_{Dihc}$ results in saturation at about 50 dB SPL (Dallos, 1985).

![Figure 5.7 Low pass filter used in the inner hair cell model](image)
The two low pass filters following the IHC nonlinearity are implemented as shown in Figure 5.7, and are implemented by the following equation:

\[ V_{out}(n) = C_1 V_{out}(n-1) + C_2 [V_{in}(n) + V_{in}(n-1)] \]  

(5.11)

where

\[
C_1 = \frac{(C - 2\pi F_c)}{(C + 2\pi F_c)} \\
C_2 = \frac{2\pi F_c}{(2\pi F_c + C)} \\
C = \frac{2}{T}
\]

The cutoff frequency \( F_c \) of these two low-pass filters is set to 1100Hz, which is in the range of estimates of the electrical membrane properties of hair cells (Russell and Sellick 1978) and results in appropriate dynamics of the hair cell model (Johnson 1980).

5.4 The Synapse

The output of the inner hair cell is a voltage which is now used as input to a model of the auditory nerve synapse. This electrical signal activates the synapse which connects the inner hair cell to the auditory nerve sending a spike train to the central auditory system. The model examined here was first proposed by Westerman and Smith in 1988, and is shown in Figure 5.8.

The model consists of two reservoirs and a global source of a neurotransmitter substance, or synaptic material. These reservoirs are connected in series and release material into the synaptic cleft by a diffusion path. The rate of diffusion (i.e. the permeability) is dependent on the input level. The global concentration is constant, independent of time and input intensity. The three permeability’s and two volumes are assumed to be functions of stimulus intensity. The local and immediate concentrations are assumed to be functions of the other six parameters and time.
Westerman and Smith's three store diffusion model in which three sources of synaptic material are connected in series and release their neurotransmitter in response to input voltage from the inner hair cell attached to the basilar membrane organ of Corti.

Figure 5.8 Model of the inner hair cell / auditory nerve synapse

The following two equations govern the concentrations in the immediate and local reservoirs:

\[
\begin{align*}
V_I \frac{\partial c_I}{\partial t} &= -p_I c_I + p_L (c_L - c_I) \\
V_L \frac{\partial c_L}{\partial t} &= -p_L (c_C - c_I) + p_o (c_o - c_L)
\end{align*}
\]  

(5.12)

These equations are required in difference form. So

\[
\begin{align*}
V_I \frac{c_I(n) - c_I(n - 1)}{T} &= -p_I (n - 1)c_I (n - 1) + p_L (n - 1)(c_L (n - 1) - c_I (n - 1)) \\
V_L \frac{c_L(n) - c_L(n - 1)}{T} &= -p_L (n - 1)(c_C (n - 1) - c_I (n - 1)) + p_G (n - 1)(c_G (n - 1) - c_L (n - 1))
\end{align*}
\]  

(5.13)

So finally

\[
\begin{align*}
C_I(n) &= C_I(n - 1) + \left[ \frac{T}{V_I(n - 1)} \right] \left\{ -p_I (n - 1)c_I (n - 1) + p_L (n - 1)(C_L (n - 1) - C_I (n - 1)) \right\} \\
C_L(n) &= C_L(n - 1) + \left[ \frac{T}{V_L(n - 1)} \right] \left\{ -p_L (n - 1)(C_L (n - 1) - C_I (n - 1)) + p_G (n - 1)(C_G - C_L (n - 1)) \right\}
\end{align*}
\]  

(5.14)

The voltage dependent permeabilities and volumes vary according to
\[ P(n) = (P_{\text{max}} - P_{\text{min}}) \frac{V_{\text{th}}(n)}{V_{\text{max}} + P_{\text{min}}} \]
\[ \text{if } (P(n) < 0) \Rightarrow P(n) = 0 \]  
(5.15)

\[ v(n) = (v_{\text{max}} - v_{\text{min}}) \frac{V_{\text{th}}(n)}{V_{\text{max}} + v_{\text{min}}} \]
\[ \text{if } (v(n) < v_{\text{min}}) \Rightarrow v(n) = v_{\text{min}} \]  
(5.16)

The output of the adaptation stage is the flow of neurotransmitter from the immediate store, and is given by

\[ S(n) = P_i(n)C_j(n) \quad \text{spikes/sec.} \]  
(5.17)

**Spike Generation**

The output \( S(n) \) represents the spikes per second transmitted by the neuron. In a discrete time implementation it can be interpreted as a probability that the neuron will fire. For example if \( s(n) = 100 \) spikes/sec, then the probability that the neuron will fire in the next bin or sampling period \( T \) is .005 (i.e. 100 x \( T \)) where \( T = 50 \mu s \). By presenting the input several hundred times and using a random number generator to determine whether or not the neuron fires during a certain period a *PostStimulus Time Histogram* can be generated.

To include refractory effects \( R(t) \) instead of \( S(t) \) is used to calculate the probability, where

\[ R(t) = S(t) - [H(t) + \theta] \]  
(5.18)

The refractory effect \( H(t) \) is given by

\[ H(t) = H_{\text{max}} \left( c_0 e^{-(t-t_i-R_A)/S_k} + c_1 e^{-(t-t_i-R_A)/S_{\theta}} \right) \quad (t-t_i) \geq R_A \]
\[ H(t) = 0 \quad (t-t_i) < R_A \]  
(5.19)
5.5 Results showing neural outputs in response to tones at high and low Q values.

Figures 5.9, 5.10 and 5.11 show the performance of the synaptic model when it is combined with the transmission line model (model 3) developed in chapter 4.

*Figure 5.9* Synaptic output in response to 80 dB tone

*Figure 5.10* Synaptic output in response to 20 dB tone

Figures 5.9 and 5.10 show the neural output from a channel of the transmission line which is tuned to 2 kHz when the input is at 80 dB and 20 dB respectively. Note that the neural discharges have increased for the higher input tone, but not by anything near 3 orders of magnitude which would be the case were the neural encoding linear. In other words the model
has compressed the input range of signals into a range which can be handled by the synaptic model, which itself has a dynamic range of about 40 dB. Note also that the way in which the model has been scaled means that the 20 dB input tone is quite near the threshold level of activity for the synapse model. Thus the synapse is more sluggish to respond in this case than for the higher input tone, in which the response is rapid. Also the initial peak in firing activity is marked for the higher tone, and the rapid adaptation whereby the synapse settles down to a resting rate of activity is more pronounced. Note also that when the stimulus is removed the firing rate initially decreases below the resting rate before recovering to the steady state firing rate of about 70 spikes/sec.

In order to keep the outputs clear and remove statistical fluctuations from the output responses (which will appear as noise when the models are used as front end processors to a neural network), the firing rate \( S(t) \) is used, and the refractory effects are ignored.

**Figure 5.11**  *Area rate curves for Modified Transmission Line model at 2 kHz*
Figure 5.11 shows the area rate curves for the modified transmission line model for a fiber tuned to 2 kHz as the input tone level is increased from 10 dB to 80 dB. Note the very sharp cutoff on the high frequency slope of the curves. This is due to the combination of all the preceding filter section in adding to the high frequency attenuation. Note also that the low frequency attenuation is not as marked since each of the preceding sections is low pass. In particular at high input signal levels the bandwidth of the model increases substantially.

From these diagrams it is to be predicted that the best performance of the model insofar as use is only being made of individual channel spectral information will be at intermediate frequencies. The other two models which have been discussed in this work produce similar results to the above when their output ranges are scaled to match the dynamic input range of the synaptic model.

Summary

This chapter has presented an overview of a number of different models for the hair cell mechanism and associated adaptation processes. The models are similar in that they all concentrate on reproducing certain characteristic behaviour on the auditory neurons, specifically the response to a tone burst, the onset and steady-state rate-intensity functions, and the adaptation times constants. They all also use some sort of neurotransmitter substance whose release is influenced by the inner hair cell voltage applied to the reservoir of material, and include mechanisms for the natural decay rate of neurotransmitter and it’s corresponding rate of generation in the stores. The model chosen to be explained in detail is the diffusion model of Westerman and Smith (1988). This model produces realistic responses to tone bursts and has accurate onset and steady state functions. It is the model used by Carney (1993) in the development of the single section gammatone filter model, and so it is suitable to compare the performances of the three models. It’s dynamic range is easily calculated, so that the output ranges of the models developed in this work can be scaled to compress whatever range of input signals is required.

In the next chapter consideration is given to the various neural network arrangements available to test the models capacity for speech recognition, and the results of a series of such tests is presented.
Chapter 6
Neural Networks and recognition results

6.1 Introduction

Neural networks take their inspiration from biological processing mechanisms in which many nonlinear ‘computational elements’ or neurons are arranged in patterns and operate in parallel. The individual computational elements or neurons or nodes connect to and from each other via connections of variable strength. The strength of these connections can be adapted during use to improve performance. Originally neural network research was beset by problems, particularly with regard to the processing power available to simulate neural networks, and the required huge processing time needed for realistic problems, and lack of suitable hardware to actually implement network models. There has been a recent resurgence in the field of artificial neural nets caused by new network topologies and algorithms (Lipmann 1987), and the growing availability of cheap but powerful processors and the belief that massive parallelism is essential for high performance speech and image recognition.

The basic premise of neural networks is that complex behaviour can be derived from the dense interconnection of many simple computational units. These computational elements are nonlinear, and are interconnected by variable weightings, by which the network as a whole can “learn” to recognise certain patterns in presented data. The simplest type of node sums $N$ weighted inputs and passes the result through a nonlinearity as shown in Figure 6.1. This diagram shows the three basic types of nonlinearity, the hard limiter, threshold logic elements and sigmoidal functions.

A neural net is defined by the network topology or architecture used to arrange it, the node characteristics (i.e. the form of the nonlinear function) and training rules. These rules specify
an initial set of weights interconnecting the various nodes, and how these weights should be altered in response to the input data. Neural nets typically provide a greater degree of robustness or fault tolerance than sequential computing since there are many more processing nodes, each with primarily local connections. Thus the networks should be more resistant to noisy input data, making them ideally suited to speech processing / recognition tasks. This contrasts with traditional statistical techniques which are not adaptive but process all training data simultaneously. Also neural net classifiers are non-parametric and do not make strong assumptions about the shapes of underlying distributions, again making them suitable for the highly nonlinear and non-Gaussian problem of speech classification.

Figure 6.1 Single node of a neural network which sums the weighted inputs and passes the result through a nonlinear function.

The history of the development of detailed mathematical models of neural networks began more than 40 years ago with McCulloch and Pitts (1943), Hebb (1949) and others. More recent work on network topologies and learning rules has been carried out by in particular Hopfield (1982, 1984, 1986) and Rumelhart and McClelland (1986).
The basic architecture of a neural network is to expand the single node shown in Figure 6.1 into a layer of nodes, in which all of the inputs connect to all the nodes, which pass the summed result through a nonlinearity. A single layer network of this type is shown in Figure 6.2. There are M output nodes for the N inputs nodes. The initial weights are typically set to random initial values between 0 and 1, as are the initial bias values ($\theta_i$) to the nonlinearities. The output of this single layer network is the result vector $Y$.

![Figure 6.2](image)

**Figure 6.2** Single layer network showing the relation between input and output vectors

In matrix notation, if the input vector is $X$, the output vector is $Y$, the weight matrix $W$, and the offset vector $\theta$, with nonlinearity $f$, then the relation between input and output vectors is

$$ Y = f(W \ast X + \theta) \quad (6.2) $$

The power of a single layer network to solve problems (i.e. to recognise certain input patterns as corresponding to a certain output) is limited to fairly straightforward, linear problems. A single layer network cannot handle any degree of nonlinearity however, the classic example being its inability to learn to recognise the exclusive or problem. In order to be able to learn to recognise nonlinear input patterns (i.e. virtually all problems of interest), another layer(s) must be added. The network shown in Figure 6.3. shows an example of such an expansion. In
In this case the middle layer is called a "hidden layer". A network of this type should be able to solve any reasonable function.

**Figure 6.3** Multiple layer network showing how intermediate results from the input to another network layer.

### The Learning Rule

A most critical aspect of the way in which neural networks function is with regard to the rules which determine how the connection strengths or weights connecting nodes together change in response to input stimuli. One of the basic techniques is the perceptron learning rule (Rosenblatt 1961), in which the network learns to generalize from the training vector and works with randomly distributed connections. This type of learning rule is thus particularly suited to classification tasks. It uses a hard limit transfer function, so that each output can take on only a binary value. It can only classify linearly separable sets of vectors, i.e. for an N-dimensional input the perceptron will only be able to separate categories of inputs if there is an N-1 dimensional hyperplane which can separate the two sets of data in N-space. Diagram 6.4 shows this concept for a 3 dimensional input vector.
3 dimensional vector space with a linearly separable set of data separated by 2-dimensional hyperplane $H$. The initial offset of the perceptron is given by $\theta$, the perpendicular distance of the plane from the origin. The perceptron 'learns' to form a representation of this plane in its internal weight matrix, and subsequently presented data should be 'classified' onto one side of the plane.

**Figure 6.4** Linearly separable data susceptible to classification by the perceptron

This restriction is due to the fact that there can only be one layer of nodes in a perceptron network, which is in turn a consequence of the fact that the perceptron learning rule can only train a single layer.

In learning to recognise data patterns, the network must first be trained with a set of training data the desired outputs of which are known. If $X$ is an input vector, $Y$ is the output and $T$ is the desired output or target vector then the weight adjustments are made as follows:

\[
W(i,j)_{\text{new}} = W(i,j)_{\text{old}} + [T(i) - Y(i)]* P(j) \\
\theta(i)_{\text{new}} = \theta(i)_{\text{old}} + [T(i) - Y(i)]
\]

(6.3)  \hspace{1cm} (6.4)

for all $i$ and $j$.

In other words, for each binary element of the output vector, if it is equal to the target vector, no change is made, if $Y(i)=0$ and $T(i)=1$, then $X(j)$ is added to $W(i,j)$ for all $j$ and if $Y(i)=1$ and $T(i)=0$ then $X(j)$ is subtracted from $W(i,j)$ for all $j$.

Each vector from the training set is presented in turn to the network. One sweep through all the training data is referred to as an epoch. There can be several tens of epochs required before the network converges to a solution, depending on how scattered the data is. After
each epoch of training the divergence between the output and target vectors is calculated. When this divergence decreases below a threshold level, or when all the input vectors have been correctly classified then training has finished and the network should be ready to classify 'new' or previously unseen data, having developed a general classification rule during training. A maximum number of training epochs should always be specified in case there is not in fact any solution for a particular set of data.

6.2 Neural networks and backpropagation

An important improvement on the basic perceptron learning rule was proposed by Widrow and Hoff (1960) in which the hard limit function was replaced by a linear transfer function, allowing the network outputs to take on a range of values rather than just ones and zeros, which in turn made it possible to develop an alternative learning rule, called Least Mean Square (LMS) in which weights and biases could be adjusted according to the magnitude of the errors and not just their presence. The Widrow-Hoff learning rule is an example of a gradient descent procedure, in which changes in the weights and biases are proportional to that weight's effect on the sum-squared error of the network. Provided that the 'learning rate' is small enough, the sum squared error is guaranteed to be minimized (Widrow and Sterns, 1985). However this learning technique is still linear, and is therefore applicable only to linear problems.

To be really useful neural nets must be capable of nonlinear function approximations. It has been shown (Rumelhart et al. 1986) that by extending the basic single layer perceptron of the last section to multiple layers, that arbitrarily complex decision regions can be formed. In particular backpropagation is a generalization of the Widrow Hoff LMS rule to multiple layer networks using nonlinear differentiable transfer functions. A set of input and target output vectors are used to train the network until it is capable of approximating a function and classifying input vectors in an appropriate way. A three layer backpropagation network which
includes (i) biases (for providing an extra free variable to help fine tune the network), (ii) at least one sigmoidal transfer function layer given by the equation

\[ f(x) = \frac{1}{1 + e^{-(x-B)}} \]  

(6.5)

and (iii) a linear output neuron layer should be capable of approximating any reasonable function. Backpropagation networks are particularly good at generalising, i.e. of classifying input vectors which have not been previously seen, which makes the technique ideal for speech recognition experiments. The following steps outline the full procedure for implementing a backpropagation network:

1. All of the initial matrix weights and biases are set to small random values. If the net is used as a classifier as it will be for speech recognition, then all output elements are set to zero except that corresponding to the class the input is from, which is set to 1. The input can be new on each epoch, or samples from a training set can be presented cyclically until the weights stabilise.

2. The aim of the backpropagation algorithm is to minimise the error over the entire set of training data. The error of the output vector \( Y=y_0, y_1, \ldots, y_{M-1} \) in response to one training example given by input vector \( X=x_0, x_1, \ldots, x_{N-1} \) with target vector \( T=t_0, t_1, \ldots, t_{N-1} \) is given by

\[ E^k(W) = (Y^k - T^k)^2 = \sum_{i=1}^{M} (Y_i^k - T_i^k)^2 \]  

(6.6)

and the total error over the entire training set is

\[ E(W) = \sum_{k} E^k(w) \]  

(6.7)

3. A gradient descent minimisation process is now carried out on the error \( E \). An approximation is used whereby each connection weight is modified following each presentation of example \( k \), using changes given by

\[ W_{ij}^{(k)} = W_{ij}^{(k-1)} - e(k) \frac{\partial E^k}{\partial W_{ij}} \]  

(6.8)
where the sensitivity of $E^k$ to each weight $W_q$ is given by

$$\frac{\delta E^k}{\delta W_q} = \frac{\delta E^k}{\delta I_i} \frac{\delta I_i}{\delta W_q}$$ (6.9)

Now if $q$ ranges over all the neurons in the layer preceding neuron $i$, then

$$\frac{\delta I_i}{\delta W_q} = \frac{\delta (\sum_p W_q O_p)}{\delta W_q} = O_j$$ (6.10)

where $O_p$ is the output of each of these neurons.

So a result for the error sensitivity is

$$\frac{\delta E^k}{\delta W_q} = \frac{\delta E^k}{\delta I_i} O_j$$ (6.11)

Replacing $\frac{\delta E^k}{\delta I_i}$ with $d_i$, we get

$$\frac{\delta E^k}{\delta W_q} = d_i O_j$$ resulting in

$$W_q(k) = W_q(k-1) - e(k) d_i O_j$$ (6.12)

4. For a neuron $i$ in the output layer, since only $S_k^i$ depends on $I_i$

$$d_i = \frac{\delta \left[ \sum_j (Y_j^k - T_j^k)^2 \right]}{\delta I_i} = 2(Y_i^k - T_i^k) \frac{\delta Y_i^k}{\delta I_i}$$ (6.13)

And since $Y_i^k = f(I_i)$

$$d_i = 2(Y_i^k - T_i^k)f'(I_i)$$ (6.14)

For the neurons in hidden layers

$$d_i = \sum_k \frac{\delta E^k}{\delta I_i} \frac{\delta I_i}{\delta S_k} = \sum_h d_h \frac{\delta S_k}{\delta I_i}$$ (6.15)

where $h$ ranges over the neurons to which neuron $i$ sends signals. But since the inputs $I_h$ to other neurons are independent of $I_i$.
\[ d_i = \sum_h d_h \frac{\delta l_h}{\delta O_i} \frac{\delta O_i}{\delta l_i} \]  
(6.16)

5. Now using an index \( p \) over neurons providing input to \( h \), these neurons are contained in the same layer as \( i \) and thus their outputs \( O_p \) are independent of \( O_i \) for \( p \neq i \), giving

\[ \frac{\delta l_h}{\delta O_i} = \frac{\delta (\sum_p W_{hp} O_p)}{\delta O_i} = W_{hi} \]  
(6.17)

6. Finally, since \( O_i = f(l_i) \), we get

\[ d_i = \sum_h d_h W_{hi} f'(l_i) \]  
(6.18)

which gives the complete rule for modifying weights when an example from the training set is presented for the \( k \)th time:

\[ W_{ij}(k) = W_{ij}(k-1) - e(k)d_i O_j \]  
(6.19)

and

\[ d_i = 2(Y_i - T_i)f'(l_i) \]  
(output layer)  
(6.20)

\[ d_i = \sum_h d_h W_{hi} f'(l_i) \]  
(hidden layer)  
(6.21)

It is a network of the backpropagation type that will be trained using outputs from the active cochlear models considered in this work. The next section gives details of the speech database used, how the cochlear preprocessing was carried out and the network architecture used.

6.3 Training the neural network with the E-set speech database

The speech database

The speech database used in these tests is the BT CONNEX speech database, which takes the following format. There are 104 speakers, 52 male and 52 female. For each speaker 24 utterances are recorded, the eight letters of the E-set of the alphabet (consisting of the letters B,C,D,E,G,P,T,V), each repeated three times. In order to train the network it is proposed that
two of the three repetitions of each letter for each speaker will be presented, thus reserving a set of samples that the network hasn’t seen before, in order to measure it’s success at generalising for unseen input vectors.

Figure 6.5 shows a sample speech file, the letter ‘p’ spoken by a male speaker. The input vector to the neural network must have a constant number of elements. Since each speech utterance will be of slightly different duration it is important to linearly timewarp the speech files to the average length of the total number of utterances. (Linear time warping is applicable here since the utterances are monosyllabic). The equation for the linear time warping of a speech signal like that in Figure 6.5 is

\[
p'(n) = p(L) + (p(L + 1) - p(L)) \times (w_n - L) \tag{6.22}
\]

where \(p'(n)\) is the linearly warped intensity contour, \(p(n)\) is the unwarped intensity contour, \(n=1,2,3,...\), \(L=\text{Integer}[w_n]\), and \(w\) is the time warping ratio (i.e. duration of the current utterance / average duration of the utterances in the set). This equation could be equally be applied to the cochlear outputs but in these tests the prewarping is first performed on the entire E-set.

![Figure 6.5 Typical speech sample from the database](image)

The speech sample like this one must, before training be linearly time-warped so that the same number of output samples will be obtained for each sample that is input to the auditory model pre-processor.
When this utterance is input to the modified transmission line model the following time/frequency plot is obtained. Notice the initial burst of activity followed by adaptation to a sustained firing rate across all the filter channels.

![Time/frequency plot](image)

**Figure 6.6** Time/frequency plot of MTL model in response to utterance of the word ‘P’.

The question now is how to reduce this amount of data into a vector that is not too large to make an attempt at training futile. In other words a way must be found to reduce the amount of data generated by the synaptic outputs (that is to reduce the dimension of the input vectors to the neural network) without losing important information. The technique chosen here is to use the Mod-Log scale Cepstral Coefficients. These features have been shown (Davis and Mermelstein 1980) to perform better than other spectral features in speech recognition systems using the Mel frequency scale. From section 2.6 the Mod-Log and Mel scale are quite similar, so in this case the same technique is used assuming that the change in scale will have no significant effect. If the 128 inner hair cell outputs of the cochlear models are represented by $S_j$, then the CC characteristics are given by the equation

$$CC_i = c_i = \sum_{j=1}^{N} S_j \cos(i(j-1/2)\pi / N)$$

(6.23)
The feature vector will consist of the 8 CC’s $c_i$ to $c_s$. This feature vector can be calculated after every 16 ms frame (or period of speech) without significantly losing information content because of the relatively slow speed at which the spectral envelope of speech changes.

The timewarping procedure which is carried out on all the speech samples prior to training determined that the average length of the training set files is 480 ms, which corresponds to 30 16 ms frames. Thus the overall feature vector for one speech sample is a 240 element CC output vector.

This is the input to the neural network. The output is an 8 element vector, since the aim of the network is to classify input vectors into one of the 8 E-set sounds. So for example for an input sound from the ‘D’ set of letters the target vector is $T = [0, 0, 1, 0, 0, 0, 0, 0]$. The input vector is fed into the network and the resulting output is calculated. This is compared with the target vector and the error, or difference between the two is calculated. The weight matrix of the network is then changed according to the back propagation algorithm. The number of hidden nodes is set to 100, since this should be enough to accurately represent the mapping from the input to the output vector spaces.

### 6.4 Recognition results

The recognition tests were run firstly with the simple inner hair cell model of Davis (1953) attached to the cochlear filters, and all the input speech files were normalised to low SPL values. At low input levels on this database it is expected that the recognition results should be at their best since the filter Q values are low and the cochlear tuning is high, and since the Cepstral Coefficient technique is exploiting the spectral, as opposed to temporal characteristics of the response. Table 6.1 shows the recognition results for the three types of cochlear filter tests, the gammatone filterbank, the alternative single channel filterbank and the Modified Transmission Line Model.
The results show that the Modified Transmission Line has the best performance of the three models. The two filterbank techniques have similar recognition rates both in this and in the subsequent tests. The sharper tuning of the MTL over the other two models which is apparent from the tuning characteristics investigated in chapters 3 and 4 are responsible for its superior performance. However the real advantage of the model becomes apparent when the ‘active’ or variable Q nature of the models are included. Tables 6.2, 6.3 and 6.4 show the recognition rates when the inputs are scaled to 30 dB, 50 dB and 80 dB input SPL’s.

### Table 6.1  Recognition results for cochlear filters using simple hair cell model with Q’s fixed high.

<table>
<thead>
<tr>
<th>Cochlear Filter</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gammatone Filterbank</td>
<td>52</td>
</tr>
<tr>
<td>Alternative Filterbank</td>
<td>54</td>
</tr>
<tr>
<td>MTL</td>
<td>65</td>
</tr>
</tbody>
</table>

### Table 6.2  Recognition results for cochlear filters using the synapse model Input SPL is 30 dB.

<table>
<thead>
<tr>
<th>Cochlear Filter</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gammatone Filterbank</td>
<td>50</td>
</tr>
<tr>
<td>Alternative Filterbank</td>
<td>49</td>
</tr>
<tr>
<td>MTL</td>
<td>58</td>
</tr>
</tbody>
</table>

### Table 6.3  Recognition results for cochlear filters using the synapse model Input SPL is 50 dB.

<table>
<thead>
<tr>
<th>Cochlear Filter</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gammatone Filterbank</td>
<td>45</td>
</tr>
<tr>
<td>Alternative Filterbank</td>
<td>46</td>
</tr>
<tr>
<td>MTL</td>
<td>53</td>
</tr>
</tbody>
</table>
These three tables indicate the essential fact about the nature of the active compression which exists in the human cochlea, i.e. that it must be sufficient to compress a wide range of input SPL’s into the smaller (40 dB) dynamic range of the neurons which are going to code the input signal and send them to higher layers of processing. It is seen here that the MTL performs better than the two single channel models across the range of input Sound Pressure Levels, but what is more important is the scale of the difference across the range of input levels. With a built in compression capability of about 20 dB the two filterbank models perform acceptably over an input range of about 60 dB, but as input levels increase they are incapable of compressing this into the dynamic range of the inner hair cell / synapse model, which when combined with the fact that the cochlear filters are broader, result in the neurons of the cochlea beginning to saturate across the basilar membrane, and leading to sharp reductions in the performance of the models in speech recognition tests. The performance of the models at a particular SPL can be changed simply by scaling the outputs of the cochlear filters with respect to the input range of the synapse, but this doesn’t change the fact that the range over which the filterbank models are effective does not change.

<table>
<thead>
<tr>
<th>Cochlear Filter</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gammatone Filterbank</td>
<td>25</td>
</tr>
<tr>
<td>Alternative Filterbank</td>
<td>28</td>
</tr>
<tr>
<td>MTL</td>
<td>44</td>
</tr>
</tbody>
</table>

Table 6.4 Recognition results for cochlear filters using synapse model. Input SPL is 80 dB.

<table>
<thead>
<tr>
<th>Cochlear Filter</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurogi without pre-emphasis</td>
<td>52.6</td>
</tr>
<tr>
<td>Kurogi with pre-emphasis</td>
<td>56.6</td>
</tr>
<tr>
<td>MLP</td>
<td>49.1</td>
</tr>
</tbody>
</table>

Table 6.5 Recognition results presented by Friel (1995) for a physiologically base neural network recogniser and MLP.
Although the performance of the models as speech recognisers is not particularly striking, the important point is their performance relative to comparable models. They compare well for example, with results presented by Friel (1995), shown in Table 6.5, which give recognition results on the same speech database for a physiological neural network model, and also for a multi-layer perceptron using a constant Q transmission line model. Thus the aim of including active compression into a model for cochlear mechanics does not necessarily mean a degradation in speech processing performance.

6.5 Summary

This chapter has presented an introduction to neural network theory, with a detailed description of the backpropagation algorithm which powers the multi-layer perceptron used in these speech recognition experiments. A description is given of the speech database which is used to train the network and thus to evaluate the models. The channel outputs in response to each speech utterance must somehow be converted into a fixed number of parameters that can be entered as the input vector to the network. A description of this process, including the calculation of the Cepstral Coefficients, is given, and finally there is a presentation of the results of a series of recognition experiments for the gammatone model, the single channel active model, and the Modified Transmission Line model (models 1, 2 and 3 from chapters 3 and 4), using a simple hair cell representation and the full synaptic model discussed in chapter 5. It is seen that the MTL model performs the best of the three, indicating that the superior frequency resolution engendered by the cascade of filter sections has a higher success rate at separating the different utterances into their constituent frequencies and therefore enabling the neural network to make a more accurate representation of the utterances in vector space.
Chapter 7

Conclusions

Three models of the active cochlea have been developed. The models were compared first of all by directly measuring their temporal and spectral characteristics, and secondly their operation was verified by means of speech recognition experiments using neural networks.

The models developed by the author were placed in perspective by introducing a number of models already in existence, both passive and active. Firstly, the basic model of Lyons (1982) was explained. Passive in its basic format, an important variation was developed by Slaney and Lyon (1983) which used a system of coupled Automatic Gain Controls to alter the overall gain of the cochlea as a function of input Sound Pressure Level, a mechanism which has parallels with the Modified Transmission Line (MTL) model (model 3) developed in this thesis. The second model described was that of Seneff (1988), a critical band model based on psychoacoustic data. The model is also active, but the total gain/damping is limited by the fact that each gain factor is implemented individually on each channel. As is seen in the development of model 2 of this thesis, this method places an inherent limit on the total amount of gain that can be achieved, since a widely varying Q factor, which would be necessary to achieve realistic compression ratios, places unacceptable delays on the latencies or fine time structures of the responses of the channels. The third model examined was that of Kates (1993), a model which manages to achieve realistic compression ratios while remaining numerically stable. This model, however, has very wide bandwidths, particularly for high level sound inputs.

Finally, Ambikairajah's model (1989) was examined, since this is the basic transmission line model which is adapted for model 3 developed by the author. The three models compared in this work are
1. Model 1 is a gammatone model as implemented in single channel form by Carney (1993). This is adapted to a parallel filterbank formation with individual channels tuned to the frequencies of the auditory spectrum according to the Mod-Log frequency scale.

2. Model 2 attempts to reproduce three aspects of the performance of model 1 which will be relevant to speech processing, namely (a) temporal response, (b) neural tuning curves, and (c) level dependency. This model attains similar behaviour to the gammatone model in terms of fine time response, and compression by means of level dependent gain control, but it has superior neural tuning curves (i.e. greater frequency resolution).

3. The problem of inadequate compression in the single channel models is addressed in model 3, a modification of the transmission line of Ambikairajah (1989). In this model the gains of the individual lowpass filters in the cascade, as well as of the finely tuned second filter are dependent on the input signal level, varying between maximum and minimum values that are computed by trial and error to produce an optimum balance between compression and overall latency of the model. While neural tuning characteristics are similar for this model and the last two, this one enables realistic compression ratios to be attained (2.5:1 over a 100 dB input range), while not excessively compromising the latencies of each channel.

A number of options for the inner hair cell and synapse models to be attached to the cochlear filter are examined. The one used is that of Westerman and Smith (1988), since it satisfies the criteria of having an accurate tonal response, adaptation time constants and onset and steady state rate intensity functions. The synapse/adaptation model has an input dynamic range of about 40 dB. Model 3 is the only one of the three models which can successfully compress 100 dB of input SPL's into this range, and therefore encode a realistic range of SPL's. The adaptation process is not strictly necessary to carry out recognition experiments to verify the cochlear filters, but is included to show how the models produce spike trains over a range of inputs.

This latter point is important, since accurate models of the auditory outputs in terms of fine time structure are an essential prerequisite for the development of theories and models as to
the higher stages of auditory processing. The purpose of the models developed here is for possible speech processing applications. In order to verify the operation of the models, and to compare them with each other, a neural network based speech recogniser is implemented. It takes as it’s inputs the Cepstral Coefficients calculated from the outputs of the cochlear filters and hair cell/synapses.

It has been seen that the third model performs the best of the three over the range of input Sound Pressure Levels, up to a level of about 65 % at high Q levels. This compares well with experiments on a constant Q cochlear model followed by a physiologically based neural network carried out by Friel (1995).

This research has thus produced a numerically stable, efficient, active cochlear model with accurate spectral and temporal characteristics. This should prove to be a good input for future research into higher levels of auditory processing, and possibly for accurate pitch detection.
REFERENCES


Appendix A

Centre Frequency Shift in Digital Filter Design

A detailed analysis of the frequency response of the digital filters described in section 4.2 shows that a frequency shift towards the lower end of the spectrum can be expected. This effect is not just due to the spatial differentiation of the displacement that takes place, but is an inherent aspect of the model. This effect is equally pronounced for the impulse invariant and bilinear design procedures.

The individual filters sections for the bilinear design are as shown in figure A.1

![Individual filter section](image)

*Figure A.1  Individual filter section*

In order to analytically calculate the displacement response of the kth section of the cascade an expression for both the pressure and displacement responses must be obtained. Then, the displacement response of filter k is equal to the product of pressure responses for filters 0 to k-1 multiplied by the pressure (pole) response of filter k.

The pressure response is obtained as follows:

\[
H_p(z) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}} \quad \Rightarrow \quad H_p(\theta) = \frac{a_0 + a_1 e^{-j\theta} + a_2 e^{-j2\theta}}{1 + b_1 e^{-j\theta} + b_2 e^{-j2\theta}} \tag{A.1}
\]
\[ H_p(\theta) = \frac{a_0 + a_1 (\cos \theta - j \sin \theta) + a_2 (\cos 2 \theta - j \sin 2 \theta)}{1 + b_1 (\cos \theta - j \sin \theta) + b_2 (\cos 2 \theta - j \sin 2 \theta)} \]  
(A.2)

\[ |H_p(\theta)| = \sqrt{(a_0 + a_1 \cos \theta + a_2 \cos 2 \theta)^2 + (a_1 \sin \theta + a_2 \sin 2 \theta)^2} \sqrt{(b_0 + b_1 \cos \theta + b_2 \cos 2 \theta)^2 + (a_1 \sin \theta + a_2 \sin 2 \theta)^2} \]  
(A.3)

Similarly, the expression for the displacement response is

\[ |H_d(\theta)| = \sqrt{(d_0 + d_1 \cos \theta + d_2 \cos 2 \theta)^2 + (d_1 \sin \theta + d_2 \sin 2 \theta)^2} \sqrt{(b_0 + b_1 \cos \theta + b_2 \cos 2 \theta)^2 + (a_1 \sin \theta + a_2 \sin 2 \theta)^2} \]  
(A.4)

Fig A.2 shows the product of the pressure responses for filters 0 to 39, the displacement response of filter 40, and the product of the two. The way in which the frequency is shifted slightly to the lower end of the spectrum is clearly visible.

**Figure A.2** Frequency responses for displacement and pressure (Bilinear)

Similar expressions can be derived for the impulse invariant design. Figure A.3 corresponds to figure Fig A.2 for the impulse invariant design. The effect is practically identical.
The actual shift in frequencies for all the filters are calculated and compared with the Mod-Log scale which was used to calculated them in the first place. The results for the bilinear and impulse invariant methods are shown in Figure A.4 and Figure A.5 respectively. These figure also shows the shift which is due to the spatial differentiation. This tends to slightly increase the centre frequency, offsetting to a certain degree the effect discussed above.

When calculating the peak frequency of the FFT outputs of the filters, the frequency resolution of the technique is obtained by the equation \((F_s/2)/(\text{no. of samples used})\). For these results the sampling frequency was reduced from 20 kHz to 8 kHz, and the number of samples was doubled from 1024 to 2048 to increase the resolution from 19 Hz to 3.9 Hz. Even so, for some of the low frequency filters there is a difference in Mod-Log centre frequencies of less than 4 Hz, so these second filters on these channels will be set to the same adjusted centre frequency.

**Figure A.3** Frequency responses for displacement and pressure (Impulse invariant).
Figure A.4  Comparison of filter peaks with the Mod-Log scale (Bilinear)

Figure A.5  Comparison of filter peaks with the Mod-Log scale (Impulse Invariant)
Appendix B

Sound Pressure Level (Miller and Arnold, 1979)

Sound, as perceived by the ear, is essentially a modulation by some vibrating object of atmospheric pressure. Under normal conditions atmospheric pressure is of the order of 15 lb/in\(^2\) or 1 bar. A variation of one millionth of the normal atmospheric pressure (1 \(\mu\)bar) is a typical stimulus for hearing, such a pressure variation being generated in normal conversation by the human voice. The minimum level of pressure variation to which the human ear is sensitive is taken in this work (to agree with the value used by Carney (1993)) to be 0.00095 \(\mu\)bar. At the upper limit of human hearing, (around 9500 \(\mu\)bars) acoustic stimuli induce pain, and possibly permanent damage to the auditory organs.

To obtain a satisfactory method of representing the range of pressure variations over which the auditory system can function we first consider the definition of power (decibels):

\[
\text{dB(power)} = 10 \log_{10}\left(\frac{\text{power} - 2}{\text{power} - 1}\right) \quad (B.1)
\]

Since acoustic power is directly related to the square of acoustic pressure this can be rewritten as

\[
\text{dB(pressure)} = 10 \log_{10}\left(\frac{(\text{pressure} - 2)^2}{(\text{pressure} - 1)^2}\right) \quad (B.2)
\]

so

\[
\text{dB(pressure)} = 20 \log_{10}\left(\frac{\text{pressure} - 2}{\text{pressure} - 1}\right) \quad (B.3)
\]

This pressure level in dB is now known as Sound Pressure Level (SPL). The lower level of human hearing is pressure-1, or 0.00095 \(\mu\)bar. Pressure-2 is the pressure variation caused by the acoustic stimulus under examination, e.g. at the upper limit of human hearing pressure-2 corresponds to 9500 mbars, so the SPL of this stimulus would be
Thus it can be seen that using the SPL system of measurement to measure acoustic stimuli, the human auditory system has a range between 0 dB (the lower threshold of hearing), and 140 dB (the upper threshold of hearing).
Appendix C

Publication:

"An Active Model of the Auditory Periphery with Realistic Temporal and Spectral Characteristics"

Amibikairajah, E., McDonagh, B.

An Active Model of the Auditory Periphery with Realistic Temporal and Spectral Characteristics

Ambikairajah, E., McDonagh, B.

Speech Research Group
Department of Electronic Engineering
Regional Technical College, Athlone
Republic of Ireland

ABSTRACT

This paper describes three auditory models which attempt to reproduce accurate temporal and spectral behaviour in a manner which is not computationally excessive. A particular problem in modelling the cochlea is to compress the input Sound Pressure Level (SPL) range into a smaller range of values (the compression ratio for a human is approximately 2.5:1), while maintaining realistic cochlear bandwidths and latencies. This paper describes a transmission line model which aims to incorporate such a compression ratio into the model, and compares its performance with that of two alternative parallel filterbank models.

INTRODUCTION

The first model presented is a modification of the transmission line model of Ambikairajah et al[1]. The filters are arranged in cascade, with the output of each filter being fed to a second filter which is tuned to the same frequency. The outputs of the second filters are sent to a peak detect mechanism which controls the Q values of the appropriate filters. This arrangement enables an almost uniform 60 dB of compression over most of the cochlear filters to be obtained. This high compression ratio is possible because the signal at any given tap has been boosted by all the preceding filter gains. The low Q values used also mean that realistic latencies are achieved (between 10 and 20 ms for the 1 kHz channel).

For comparative purposes another recently proposed model is considered, that of Carney (2), in which a single section of the basilar membrane is modelled by a gammatone filter. The bandwidth is controlled by a level-dependent non-linear saturating feedback, which simulates the activity of the Outer Hair Cells, and the parameters are chosen to most closely match the responses of cat auditory nerve fibers. In this paper, a number of these gammatone filters are arranged in a parallel filterbank formation, with each filter tuned to a different frequency of the auditory spectrum. Using the parameter values specified by Carney[2] for all filters, the tuning curves resulting from this model seem to be excessively broad as compared with those of the first model, and the signal compression is less than that achievable with the modified transmission line technique.

As a third model, an alternative parallel filterbank is presented, which reduces the computational effort involved, while maintaining the compression ratio and achieving greater frequency selectivity. In this model each section of the membrane consists of a second order bandpass filter. The output of this filter is fed to a second filter tuned to the same centre frequency. The second filter has a variable Q, which is dependent on the signal level. It is controlled by a capacitor hair cell model. By varying the value of Q, the broadening of the cochlear bandwidth and 20 dB of compression can be realised in a more efficient fashion to that of Carney.

For the three models the output of the cochlear filters are input to the synapse model proposed by Carney [2]. This consists of an inner hair cell, (a saturating nonlinearity followed by two lowpass filters), and an inner hair cell-auditory nerve synapse. The three models are compared under the three main criteria of (a) temporal response, (b) neural tuning curves, (c) level dependence.

The design procedure for the implementation of the three models is described. All three models operate at a sampling frequency of 20 kHz and cover the frequency spectrum from 75 Hz to 3.5 kHz. The principal intended use for these models is as front-end processors for speech recognition systems.
The transmission line model of the cochlea of Ambikairajah \textit{et al} \cite{1}, uses 128 cascaded filters to represent adjacent sections of the basilar membrane. Each filter is tuned to an individual frequency of the auditory spectrum, which ranges from 3.5 kHz to 75 Hz. As a pressure signal propagates along the cascade high frequency components are filtered out by the 2nd order lowpass filters. The pressure transfer function for a single section of the membrane, and it's digital filter realisation are given in Figure 1.

\begin{align*}
\frac{V_1}{V_0} &= \frac{1 - B_{Z1} - \omega^2}{1 - B_{Z1} - B_{P1} - \omega^2} \frac{\omega^2 - B_{P1} - \omega^2}{\omega^2 - B_{Z1} - \omega^2} \\
Q_{Z1} &= \frac{1}{B_{Z1}} \quad Q_{P1} = \frac{1}{B_{P1}}. 
\end{align*}

Figure 1. A single section of the digital filter model of the basilar membrane.

In the modified transmission line model, a second filter is used at each tap to enhance the gain and selectivity of the model. This filter is tuned to the same frequency as the corresponding filter in the cascade structure (Figure 2). Displacement outputs from the first filters undergo two spatial differentiation's to accentuate the tuning characteristics before being input to the second filters. This differentiation represents the mechanical coupling of adjacent sections of the membrane. The transfer function for the second filter is the same as the pressure transfer function of the cascaded filters.

The second filter output is to be used to control the filter bandwidths and gain. That is, at low signal levels the gain is high and the bandwidth is low, and at high signal levels the gain is low and the bandwidth is broader. This dynamic in the gain and bandwidth of the filters is obtained by adjusting the Q's of each filter (Kates\cite{3}) as a function of the peak output from the second filters. The maximum Q values ($Q_{P1}$) of the cascaded filters vary from 1.2 to 8.0 from the low to high frequency ends, on a linear cochlear distance scale. The relationship between the zero and pole parameters are: $\omega_z(i) = 1.15\omega_z(i-2)$ ($i$ = filter number, $\omega_z$ = zero resonant frequency, $\omega_p$ = pole resonant frequency), and $Q_{z1} = 1.50Q_{z1}$. For the second filter the maximum Q values ($Q_{P1}$) are also frequency dependent, determined by the relationship $Q_{P2} = 1.5(1+f)$ with $f$ in kHz, and the zero/pole parameters vary according to $\omega_z = \omega_p / 2$ and $Q_{z2} = 2.0Q_{z2}$.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{model1.png}
\caption{Transmission Line Model (Model 1)}
\end{figure}
The sustained rate has a dynamic range of about 40 dB. It is because of this physiological limitation in the ability of the auditory nerve to encode signals over a large dynamic level that the cochlea must compress the enormous variation of sound inputs into a more manageable range. Since the transmission line model discussed in the preceding section has a compression ratio of 2.5:1 over an input range of 100 dB, the auditory synapse model is able to successfully encode this range of signal inputs.

The feedback control mechanism works as follows: the outputs of each of the second filters are sent to a peak detect mechanism. This detects the peak output over a span of 8 filters either side of the one whose Q is to be controlled. This peak output is then used to determine the ratio of peak output to maximum output (r), which in turn controls the Q setting of each filter. The cascaded filter Q’s are given by \( Q_{f_r} = (0.3 + 0.7r) \times Q_{\text{max}} \) and the second filter Q’s are given by \( Q_{f_\theta} = (0.1 - 0.9r) \times Q_{\text{max}} \). Filter outputs should lie between 40 and 80 dB to ensure that the 40 dB range of filter outputs matches the 40 dB dynamic range of the inner hair cell/synapse model. This is why the frequency responses are scaled to peak at -20 dB at low Q.

Figure 3 shows the frequency responses of every 6th filter when all the Q’s (\( Q_{r} \) and \( Q_{\theta} \)) are fixed in both the low (Fig. 3a) and high (Fig. 3b) states. A scaling factor is used to give all the responses a maximum gain of -20 dB at low Q. In the high Q state, both the low and high frequency responses do not have a full 60 dB gain. This is due to the fact that for the high frequency filters there are not enough preceding filters for the gain to have accumulated to 60 dB, while for the low frequency case it is because the overall gain at these frequencies does not sum to 60 dB.

Figure 4 shows the sustained firing rate of the synapse for an intermediate frequency channel. The firing rate is computed (as in [2]) from the average output of the synapse over the duration of the input tone. The first 5 ms of the response is discarded so as not to include the initial burst of activity. The firing rate is almost constant at low signal input levels. Then at about 40 dB SPL, the firing rate increases steadily until it eventually saturates.

The sustained rate has a dynamic range of about 40 dB. It is because of this physiological limitation in the ability of the auditory nerve to encode signals over a large dynamic level that the cochlea must compress the enormous variation of sound inputs into a more manageable range. Since the transmission line model discussed in the preceding section has a compression ratio of 2.5:1 over an input range of 100 dB, the auditory synapse model is able to successfully encode this range of signal inputs.

Carney’s hair cell model

For the purpose of comparing the three cochlear models the same inner hair cell and synapse model is used in all three cases.

Figure 4 shows the sustained firing rate of the synapse for an intermediate frequency channel. The firing rate is computed (as in [2]) from the average output of the synapse over the duration of the input tone. The first 5 ms of the response is discarded so as not to include the initial burst of activity. The firing rate is almost constant at low signal input levels. Then at about 40 dB SPL, the firing rate increases steadily until it eventually saturates.

The sustained rate has a dynamic range of about 40 dB. It is because of this physiological limitation in the ability of the auditory nerve to encode signals over a large dynamic level that the cochlea must compress the enormous variation of sound inputs into a more manageable range. Since the transmission line model discussed in the preceding section has a compression ratio of 2.5:1 over an input range of 100 dB, the auditory synapse model is able to successfully encode this range of signal inputs.
MODEL 2

\[ g(t) = \begin{cases} 0 & t < \alpha \\ \frac{(t-\alpha)}{\tau} \exp(-t/\tau) \cos(\omega_c \, (t-\alpha)) & t \geq \alpha \end{cases} \]

In this paper 128 such gammatone filters are arranged in a filterbank formation. Figure 6 shows the frequency responses of every 6th filter in the filterbank. These responses are obtained by taking an FFT of the impulse response of each filter with the parameters set to those specified by Carney.

MODEL 3

For the purposes of speech recognition experiments the essential features of latency and cochlear tuning can be obtained using a much simpler, alternative parallel filterbank model, which is very much concerned with physiological accuracy with regard to the feedback control mechanism and the cochlear bandwidth.
This model uses a bandpass filter followed by a second-order lowpass filter, both of which are tuned to the same characteristic frequency. The bandpass filter output is used to control the Q factor of the lowpass filter, which influences the gain and bandwidth characteristics of the filter (Figure 8). The feedback control mechanism consists of a simple capacitor outer hair cell model, the output of which is integrated over a short period to get the energy. This integrated value is compared with the pre-computed maximum and minimum values of OHC energy and used to change the lowpass filter coefficients.

Figure 9 shows the frequency response of the filterbank at high and low values of Q. There is a 20 dB difference in maximum gain between the two states.

RESULTS

Figure 10 shows a comparison between the temporal responses of the three models in response to an input tone at CF at two different SPL's. The time taken for the synapse output to decay to its resting rate after the signal has been removed indicates the latencies of the models. The three models all give reasonable latencies of less than 20 ms for an 80 dB signal, when the filter Q's are low. The effect of the superior compressive abilities of model 1 are evident in the synaptic response to a low level input.
Figure 11 shows the rate level curves for the three models. The synaptic output is plotted for a particular frequency channel, in response to tones at CF at varying SPL's. Models 2 and 3 both show a dynamic range of approximately 50 dB. Due to the high compression ratio in the case of model 1, response can take place over a much larger range of inputs (approx. 90 dB).

CONCLUSIONS

The Transmission Line model (Model 1) is evidently a superior cochlear model under the three criteria examined above. A realistic compression ratio is unattainable using a parallel filterbank formation (Models 2 and 3) of independent filters. The cascade structure model 1 also ensures a superior frequency selectivity in the channels of the auditory nerve, however the latencies of Model 1 are slightly longer than for the other two models.

ACKNOWLEDGEMENTS

This research has been carried with funding provided by EOLAS, the Irish Science and Technology Agency, to Regional Technical College, Athlone under the Applied Research Program 1993/94.

REFERENCES

