ANALOGUE VLSI BUILDING BLOCKS FOR AN ELECTRONIC AUDITORY PATHWAY

PAR

Floris André VAN SCHAIK

Ingenieur elektrotechniek, Universiteit Twente, Enschede, Pays-Bas
de nationalité hollandaise

to Yolanda

Acknowledgements

First of all, I wish to express my gratitude to my advisor, Professor Eric Vittoz, for the things he taught me over the years, his insights and advice. I'm especially grateful to him for having enough confidence in me to allow me to work on the subject presented in this thesis, even when it was clear that this would not lead to a standard electrical engineering thesis.

I'm also deeply grateful to Professor Ray Meddis without whom this work would have been impossible. I thank him for patiently answering my emails with questions (there were a lot of those), for his collaboration on parts of this work, for his insights, and his friendship.

I would also like to thank Professor Wulfram Gerstner for helpful discussions on spike-based computation, for agreeing to be one of my referees, and for his constructive criticism on the draft version of my thesis.

Furthermore, I would like to thank Dr. Xavier Arreguit, for agreeing to be one of my referees and for his enthusiasm for neuromorphic engineering, but mainly for being my mentor at CSEM before I started my Ph.D. research; he helped me to become a better engineer.

I thank Eric Fragnière for being the other member of the "Analogue VLSI for Perception" group. Sharing the same office and working together for four years is a bit like being married, although I must say that we quarrelled surprisingly little. Without Eric to test my ideas on and without his feedback I would not have been able to do this work.

I am grateful to Professor Jean-Daniel Nicoud for the MANTRA environment and to my colleagues at MANTRA for making it a pleasant environment to work in (and have coffee breaks in). Special thanks goes to Alix Herrmann for proof reading my thesis and for helping me clarify some of the more difficult passages.

I wish to thank Jean Hennebert and Professor Martin Hassler for their insights on the subject of speech recognition. It helped me recognise at an early stage that we are still far from bio-inspired speech recognition.

I am deeply grateful to my parents for their support and love.

Résumé

Cette thèse présente un aperçu de mon travail effectuée ces quatre dernières années sur le développement de modules de construction électroniques analogiques intégrés pour le système auditif et leurs applications à certains modèles de traitement d'information dans le système auditif central.

L'anatomie et la physiologie de l'oreille humaine est présentée et est décomposée en trois éléments-clé: le filtrage 'passe-bande' effectué par la *membrane basilaire*, la transduction en signal nerveux par les *cellules ciliées internes* et la contre-réaction mécanique par les *cellules ciliées externes*. Un modèle électronique est présenté pour les deux premiers éléments et des résultats de mesures sont montrés afin de comparer leur comportement avec celui de leur équivalent biologique.

Le reste du système auditif consiste en plusieurs groupes de différents *neurones impulsionels*. Etant donné que la majeure partie du traitement du signal dans le système auditif central est effectué par ces différents types de neurones, un bon modèle du neurone impulsionel est essentiel. L'électro-physiologie et l'anatomie nécessaire à la compréhension du fonctionnement de base de ces neurones est présentée.

Dans le but de modéliser de grands groupes de neurones, le circuit modélisant un neurone doit être suffisamment petit. Je propose un circuit petit et simple, permettant cependant d'émule les différents types de neurones présents dans le noyau cochléaire, ainsi que montré par les mesures effectuées sur le circuit.

Grâce à ces modules de constructions électroniques, je peux envisager la modélisation d'architectures neurones capable d'extraire en temps réel différentes caractéristiques du signal auditif. Bien que peu de ces architectures aient été identifiées dans le cerveau, une fois que les types de neurones y participant ainsi que leur interconnection sont connues, il est relativement facile d'en réaliser le modèle analogique intégré. J'en présente deux exemples, basées tous deux sur la détection de périodicité, qui est un mode de traitement pour lequel les neurones impulsionels semblent particulièrement approprié.

Le premier exemple exploite l'activité synchrone de fibres du nerf auditif venant de deux positions différentes le long de la membrane basilaire afin d'obtenir une grande sélectivité fréquencielle et une représentation du signal

auditif indépendante de son intensité. Ce modèle est complètement hypothétique, n'ayant pas été découvert dans le cerveau, mais il donne une excellente introduction à la puissance du traitement effectué par les neurones impulsionel pour la détection de périodicité.

Le second exemple est un modèle de la sensibilité à la modulation d'amplitude dans le colliculus inférieur. Ce modèle peut être utilisé par exemple pour extraire la fréquence de modulation d'un signal modulé en amplitude ou pour extraire la fréquence fondamentale d'un complexe harmonique. Ce type de signaux sont particulièrement intéressant parce que de nombreux sons naturels tombent dans cette catégorie. Ce modèle a des fondations biologiques plus solides, des neurones réagissant similairement à la modulation d'amplitude ayant été mis en évidence dans le colliculus inférieur.

Abstract

This thesis gives an overview of my work over the last four years on the development of analogue electronic building blocks for the auditory pathway, and their application to some models of processing in the auditory brainstem.

The anatomy and physiology of the human ear is presented, and is decomposed into three key elements, i.e., the *basilar membrane* 'band-pass' filters, the transduction into a neural signal performed by the *inner hair cells*, and the mechanical feedback introduced by the *outer hair cells*. An electronic model for the first two of these elements is presented and measurement results are shown to compare these circuits with their biological counterparts.

The remaining part of the human auditory pathway consists of several groups of different types of *spiking neurons*. Since the main part of signal processing in the auditory pathway is performed by these different types of spiking neurons, a good spiking neuron model is essential. The electrophysiology and anatomy needed to understand the basics of the spiking behaviour of these neurons is presented.

If we want to model large groups of spiking neurons, the neuron circuit also needs to be small. I propose a circuit that is simple and small, yet capable of emulating several types of neurons found in the Cochlear Nucleus, as shown by chip measurements.

With these electronic building blocks I can start to model neural architectures in the brain that extract certain signal characteristics, and these models will operate in real time. Although only few of such architectures have been identified to date, once the types of neurons in these brain circuits, and their inter-connectivity is known, it is then fairly straightforward to create an analogue VLSI model. I present two examples, both based on periodicity detection, since this is a domain where spike based operation seems especially useful.

The first example uses synchronised activity on auditory nerve fibres from two positions along the basilar membrane to obtain a high frequency selectivity and a representation of the sound which is independent of intensity. This model is completely hypothetical, since no evidence has been found in the brain for its

existence, but it provides an excellent introduction into the periodicity detecting power of spike based computation.

The second example is a model of amplitude modulation sensitivity in the inferior colliculus. This model can be used for example to extract the modulation frequency of an amplitude modulated sound, or to extract the fundamental frequency of a harmonic complex. These sounds are of special interest because many natural sounds such as animal calls or speech fall into this category. This model has a much stronger biological basis, since neurons have been found in the inferior colliculus that react to amplitude modulation in a similar way as the model.

Contents

1. Introduction

The work in this thesis was inspired by the ease with which animals perform perceptive tasks. Perception is the result of neural processing that extracts from the sensory input the characteristics needed to form an internal representation of the environment and to interact with this environment. Not only does the nervous system perform these tasks well (enough to allow the animal to survive), it does so using imprecise and unreliable processing elements (neurons) and sensors. It is my conviction that for these perceptive tasks, natural selection has developed processing strategies that are far more effective than those found in man-made systems today.

Analogue integrated circuits and most small, cheap sensors also suffer from imprecision problems. We may hope that by using similar processing strategies as biological organisms, it might be possible to overcome the imprecision of the analogue building blocks and to perform perceptive tasks on silicon just as effectively. The only way to verify this hypothesis is by building analogue electronic systems that function according to these neural processing strategies.

Neural computation in the brain is performed by an immense variety of neurons, whose behaviours are diverse and complex; to fully describe them requires a large number of parameters. However, we might not need to replicate biological neurons in full detail in order to build effective neural systems. On the other hand, because of natural evolution it is quite possible that many details of the behaviour of a biological neuron are not negligible. We will therefore have to determine the right amount of detail. One approach is to build a functional model that embodies as many details as possible; once it is built, it should be possible through experimentation to distil the amount of detail needed for each specific function and thus derive a simpler model.

A second reason for building electronic models of neural processing is that these models will allow the study of neural processing of the brain without the need for living neural tissue and without the need to simultaneously probe hundreds or thousands of neurons in the brain. Once we have neuron models with the same characteristics as biological neurons, together with knowledge of their inter-connections, we may build a hardware model that allows us to study

the collective behaviour of such a neural network and the individual behaviour of the single neurons in the network.

In principle, we can do the same using computer models. However, as our understanding of the actual processes in the brain increases, these models are becoming more and more detailed. Consequently the computer models are becoming more and more computationally intensive and memory demanding. This is a serious problem, which threatens to take the simulation of these models beyond the range of even the most powerful digital computers. Electronic building blocks, however, should not suffer from this problem because their parallel architectures permit them to operate in real time.

In any case we will not be able to build hardware implementations of neural processing using just one neuron, albeit carefully modelled. We will need a large number of neurons in parallel to be able to model anything substantial. This leads to a trade-off between the size of the single-neuron circuit and the amount of detail incorporated in it. On the other hand, even if each chip only models a few hundred neurons, multiple chips may be used in parallel. There is thus no obvious limit to the number of neurons in an analogue VLSI model.

Most computation in the brain is performed using spikes, i.e., small voltage pulses of a fixed width and fixed height. *A priori*, spike-based computation seems especially efficient when we want to determine whether two signals are more or less synchronous; a simple coincidence detection operation will do the job. To study this property, we shall have to look at neural systems that make use of the temporal relations between signals. The auditory system is an excellent candidate, since its inputs are just two one-dimensional variables (air pressure at both ears) that vary with time.

Any neural system in the auditory pathway that we try to model will include input from at least one ear. Therefore, we should start by developing an electronic model of the ear. The main signal-processing structure in the ear is the cochlea. It contains the basilar membrane, which mechanically filters the sound input, and hair cells, which transduce basilar membrane motion into neural signals.

Building these elements as separate chips will give us the flexibility to create different models using the same building blocks. Once a basic library of these building blocks has been developed we can connect these chips together to build models of different parts of the auditory pathway. In order to do this, however, we must be able to control the important parameters of the building blocks off-chip; we shall also have to include one or more inter-chip communication possibilities that make flexible connections possible.

In the first part of this thesis I will present some of these building blocks along with the biology needed to understand them. In the second part I will then describe two neural systems implemented with these building blocks.

Chapter 2 will briefly present the anatomy and function of the human ear. Special attention will be given to the key elements of the cochlea, i.e., the mechanical filtering of the basilar membrane and the mechanical to neural transduction by the inner hair cells. The 3 electronic models for each of these elements will then be presented in Chapter.

In Chapter 4 we shall look at the spike generation process in the biological nerve cell and discuss the influence of different anatomical parameters on the behaviour of the cell. Chapter 5 will present an electronic model for this important building block of the auditory pathway. In different settings, this circuit can be used to model different types of neurons. The output of the circuit in response to tone-bursts will be compared with the responses of the different types of neurons in the cochlear nucleus, which is the first site of innervation of the auditory nerve.

Chapter 5 will furthermore present several auxiliary circuits used for the interaction between neurons. One may think of these circuits as the electronic equivalent of dendrites, axons and synapses.

Chapter 6 will present a first example of application of the building blocks to the modelling of neural systems. It describes a system that implements a model of the extraction of periodicity information from the output of the cochlear filters.

Chapter 7 describes a second example, which models the sensitivity of neurons in the cochlear nucleus and inferior colliculus to the envelope periodicity of amplitude-modulated sounds and harmonic complexes. This system has been implemented using the separate building blocks. Furthermore, a new dedicated version, which integrates several of the building blocks on a single chip, will also be presented.

Finally, the conclusions will be presented in Chapter 8, which will sum up the main strengths and weaknesses of this work, and will give some recommendations for future research.

.

2. The Human Ear

2.1 Introduction

This chapter will describe in some detail the anatomy of the outer, middle and inner ear. Furthermore the electrophysiology of the cochlea will be discussed. The basic functions of the cochlea, which is the most important element from a modelling point of view, will be discussed in section 2.4.

Most information on the anatomy and physiology of the ear is taken from Dallos (1984), Evans (1982), Kessel and Kardon (1979), and Pickles (1988) and is based on measurements and observations of different animals. Nevertheless, the goal of this chapter is to understand the human ear; most of the elements shown by the animal data also hold true for humans.

2.2 Outer and Middle Ear

The outer ear consists of the pinna, the concha, and the ear canal, which leads to the eardrum (Figure 2.1). The main function of the outer ear is to protect the delicate eardrum from the external environment. Furthermore, the outer ear performs two signal processing functions. Firstly, for some frequencies it increases sound pressure at the eardrum due to resonance in its cavity. Secondly, the outer ear provides cues for sound localisation by altering the spectrum of the sound depending on the position of the source.

The middle ear is an air-filled cavity within the bone of the skull. The ossicular chain is suspended in this cavity by small muscles. In humans the chain consists of three bones: the hammer or malleus, the anvil or incus, and the stirrup or stapes. The malleus is firmly attached to the eardrum, the incus comparatively rigidly to the malleus and the stapes is attached to the oval window. In this manner, the ossicular chain ensures that sound which reaches the eardrum is only transmitted to the oval window. Without the ossicular chain, sound would reach both the oval window and the round window of the cochlea at the same time, and the fluid in the cochlea would hardly move. Furthermore, together with the eardrum and the oval window, the ossicles provide acoustic impedance matching. This increases the efficiency with which air-borne sound energy is transferred to movement of the fluid-and-membrane structure of the

cochlea. The main mechanism behind this impedance matching is the concentration of the forces collected over the eardrum to the much smaller area of the stapes footplate. Other factors are a small lever action due to the different length of the malleus and incus, and the conical shape of the eardrum which makes the eardrum buckle as it moves in and out so that the eardrum moves more than the arm of the malleus.
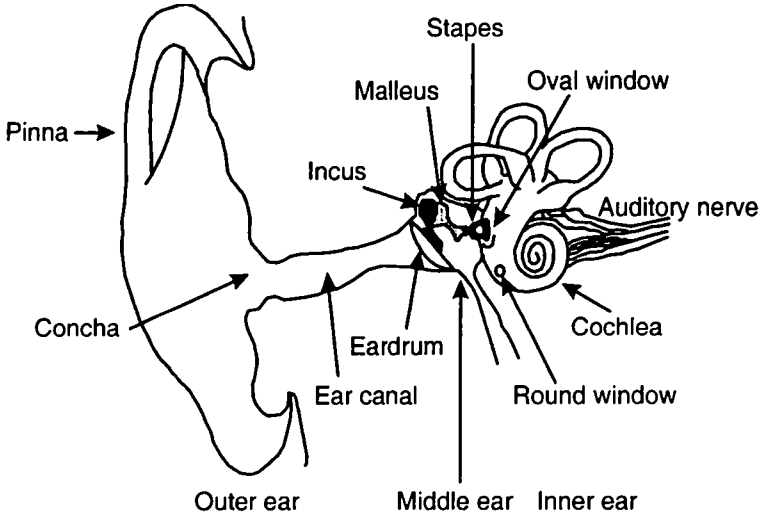

Figure 2.1 The human outer, middle and inner ear.

The ossicular chain has a certain resistance to displacement. This reduces the efficiency of the transmission, mostly for low frequency sounds since displacement is inversely proportional to frequency at constant sound pressure. At high frequencies the transmission is reduced by many factors, including the mass of the ossicles and less efficient modes of vibration of the structures. The pressure gain of the middle ear as a function of frequency therefore has a weak band-pass characteristic, with maximum gain around 1 kHz.

Two small muscles (not shown above) are attached to the ossicles. The tensor tympani is attached to the malleus near the tympanic membrane and the stapedius muscle is attached to the stapes. Both muscles receive descending connections from the brain. When the muscles contract, the resistance to displacement of the ossicular chain is increased. The middle ear muscles therefore influence the gain at frequencies below 1 kHz most, since the displacement is relatively large at these frequencies. As we shall see later, the low frequency sounds travel the whole length of the cochlea, and therefore can potentially damage any part of the cochlea. The contraction of the middle ear muscles for loud sounds thus protects the inner ear from noise damage to some

extend. However, it does not protect against impulsive noise, because the reflex is quite slow. These muscles also contract just before we start talking or swallowing, reducing the ear's response to the low frequency part of these internal noises which could otherwise mask high frequency external sounds.
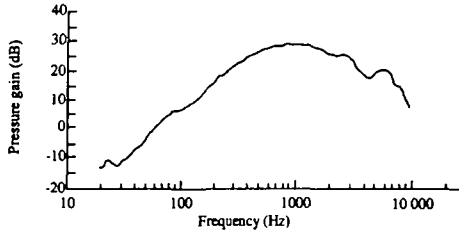

Figure 2.2 Transfer function of the middle ear of the cat. Adapted from Nedzelnitsky (1980).

In general, little attention is given to the modelling of the outer and middle ear filtering function. The combined effect of outer and middle ear is that of a band-pass filter with a slight gain between 1 and 6 kHz, a 12 dB per octave cut-off slope above 6 kHz, and a 6 dB per octave slope below 1 kHz, which may be steepened by the auditory reflex. Except for this last effect, the outer and middle ear could be modelled using a simple filter at the input of the cochlea, but for the purpose of the experiments in this thesis its effect would be negligible. This filter has therefore not been developed.

2.3 Anatomy of the Inner Ear

This section will present a relatively detailed description of the anatomy of the inner ear. The information in this section will give a reader who is not familiar with these details a better background to understand the functioning of the inner ear. However, only the elements presented in section 2.4 on the functioning of the cochlea will really be important for the creation of electronic models.

The inner ear is a fluid-filled bony structure embedded deep in the temporal bone of the skull and contains the semicircular canals, the vestibule, and the cochlea. The stapes presses on the oval window, an opening in the vestibule. Vibration of the stapes causes pressure waves to travel in the fluid inside the vestibule and the cochlea. The round window provides pressure relief for the incompressible cochlear fluid. The cochlea itself is a coiled structure and contains the organ of Corti. In humans the cochlear canal is approximately 35 mm long, has 2.5 turns and is wider at the base (close to the oval window) than at the apex (far from the oval window). Figure 2.3a shows how the cochlea would look if it could be uncoiled. Figure 2.3b shows a cross-section through the uncoiled cochlea.
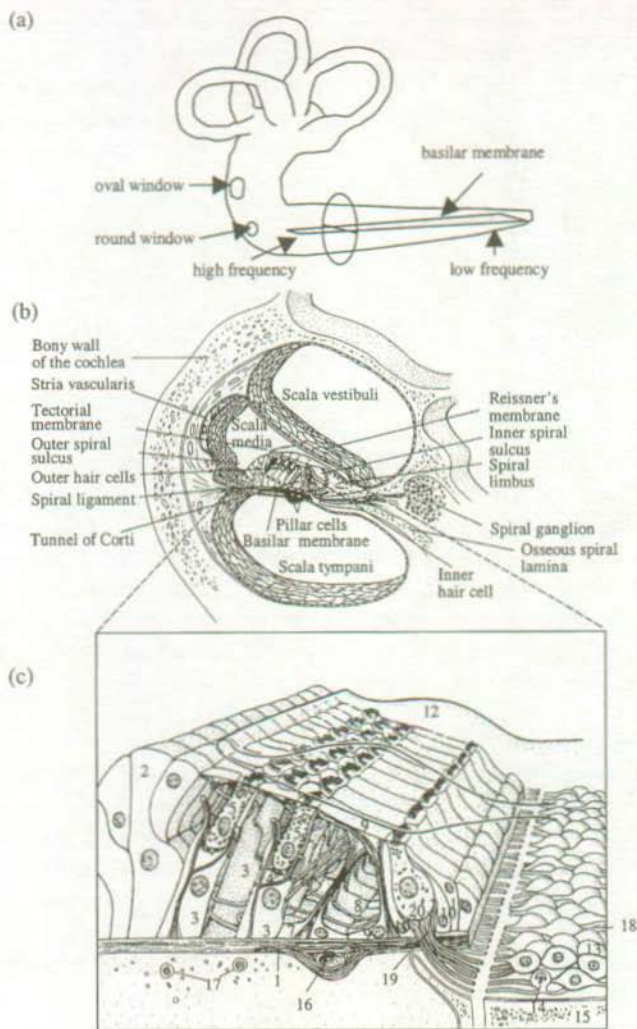
(a)

basilar membrane

oval window

round window

high frequency   low frequency

(b)

Bony wall
of the cochlea
Stria vascularis
Tectorial
membrane
Outer spiral
sulcus
Outer hair cells
Spiral ligament

Tunnel of Corti

Scala vestibuli

Scala
media

Pillar cells
Basilar membrane
Scala tympani

Reissner's
membrane
Inner spiral
sulcus
Spiral
limbus

Spiral ganglion
Osseous spiral
lamina
Inner
hair cell

(c)

Figure 2.3 a) The cochlea uncoiled; b) Cross-section through the uncoiled cochlea; c) The Organ of Corti. The first outer hair cell (the most basal one) of the middle row is removed so that three-dimensional aspects of the relationship between supporting cells and hair cells can be seen. 1, Basilar membrane; 2, Hensen's cells; 3, Deiters' cells (outer phalangeal cells); 4, endings of spiral afferent fibres on outer hair cells; 5, outer hair cells; 6, outer spiral fibres; 7, outer pillar cells; 8, tunnel of Corti; 9, inner pillar cells; 10, inner phalangeal cells; 11, border cell; 12, tectorial membrane; 13, type I spiral ganglion cell; 14, type II spiral ganglion cell; 15, bony spiral lamina; 16, spiral blood vessel (found only in the base of the cochlea); 17 cells of the tympanic lamina; 18, auditory nerve fibres; 19 radial fibre; 20 inner hair cell. Adapted from Kelly (1991).

The cochlea is divided into the scalae vestibuli, media and tympani by Reissner's membrane and the basilar membrane together with the osseous spiral lamina. The osseous spiral lamina is wide (in the direction of the basilar membrane) at the base, about 77% of the diameter of the cochlea, and narrows gradually to only 40% near the apex. A hole in the spiral lamina at the apex, called the helicotrema, joins the scalae vestibuli and tympani. Both the basilar membrane and Reissner's membrane stop before the helicotrema, thereby terminating the scala media. Whereas the spiral lamina narrows, the basilar membrane widens from base (about 100 μm) to apex (about 500 μm). As the basilar membrane widens, its stiffness decreases more than 100-fold in an approximately exponential fashion. Reissner's membrane provides ionic isolation between the scalae media and vestibuli but serves no mechanical purpose in the filter function of the cochlea. The fluid contained in the scalae vestibuli and tympani, called perilymph, is high in sodium content and low in potassium content, similar to the extracellular fluids in the brain, and is at about the same potential as the surrounding bone. The scala media is filled with endolymph, which is low in sodium but rich in potassium and has a potential of about +80mV with respect to the surrounding bone. The difference in ionic concentration between the endolymph and perilymph is maintained by the dense capillary network called the stria vascularis. The potential difference thus created acts as a low-noise power supply for the hair cells in the organ of Corti. Since these hair cells are sensitive to very small movements, they must be isolated from the noise introduced by the circulatory system. There is a small blood vessel beneath the basilar membrane, as shown in Figure 2.3c, but no capillaries extend into the organ of Corti.

The organ of Corti resides on top of the basilar membrane in most mammals and contains the actual mechano-electrical transducers: the hair cells. About 3000 inner hair cells form a single row, while about 9000 outer hair cells are arranged into three to five rows. The hair cells are spaced about 10 μm apart and are rigidly attached to the basilar membrane by the supporting Deiters' cells and the pillar cells. Deiters' cells have processes that extend upward to hold the tops of the outer hair cells, resulting in the rigid upper surface of the organ of Corti, called the reticular lamina. The inner pillar cells rest on the basilar membrane just above the end of the osseous spiral lamina, which is the point around which the complete organ of Corti pivots. Hair cells are so named because of the stereocilia, or hairs, that protrude from their smooth and stiff apical surface called the cuticular plate. This can clearly be seen in Figure 2.4, which is a scanning electron micrograph of the top of the organ of Corti with the tectorial membrane removed.
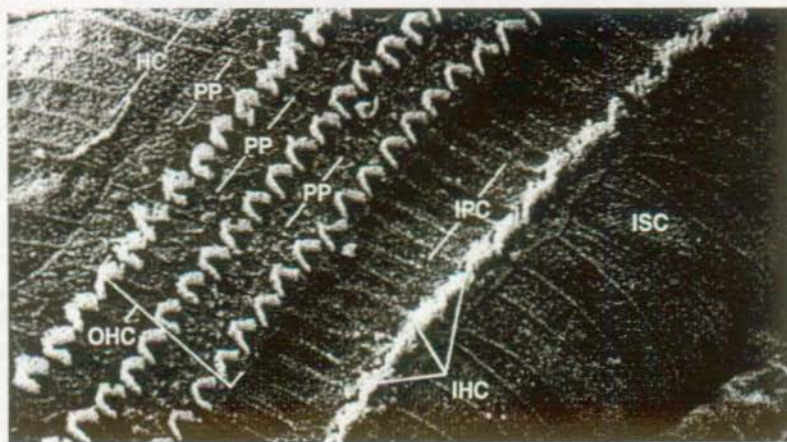
Figure 2.4 Scanning electron micrograph of the organ of Corti after removal of the tectorial membrane. The single row of inner hair cells (IHC) contains stereocilia that are arranged linearly. The three rows of outer hair cells (OHC) contain stereocilia that are arranged in V or W configurations. The surfaces of a number of other cells may also be distinguished. These include the inner spiral sulcus cells (ISC), heads of the inner pillar cells (IPC), phalangeal processes (PP) of Deiters' cells, and Hensen's cells (HC). Adapted from Kelly (1991).

The stereocilia are formed by closely packed actin filaments which make them rigid. Just above the cuticular plate the stereocilia taper, which allows them to pivot around this point. Figure 2.5 shows a cross-section through three rows of stereocilia on a hair cell. This clearly shows that the stereocilia are linked between rows by horizontal links and tip links. It has also been shown that stereocilia within a row are linked together by horizontal links. Tip links however only run from the tip of a stereocilium to the nearest taller stereocilium.

There are clear morphological differences between the inner and outer hair cells, as shown in Figure 2.6. The stereocilia of the inner hair cells are arranged in shallow curves, and in V or W formations for the outer hair cells. The tallest cilia of the outer hair cells are attached to the tectorial membrane, whereas the cilia of the inner hair cell are not. The inner hair cells are flask-shaped, flexible, and are completely surrounded by the inner phalangeal cells. The outer hair cells are cylindrical in shape, stiff, and only attached to the supporting structure at their tops and bottoms, so that most of the cell is free to move. The outer hair cells contain fine tensile filaments that wrap around the cell body to form a kind of skeleton structure that provides, together with the turgor pressure of the cell's cytoplasm, a mechanism to change the length of the outer hair (Ashmore, 1987; Crawford and Fettiplace, 1985).
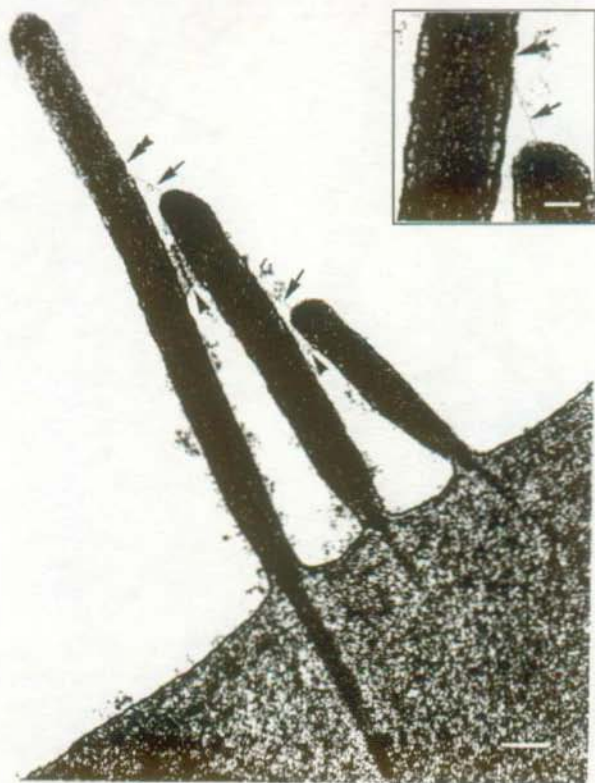
Figure 2.5 Cross-section through three rows of stereocilia on a single guineapig outer hair cell. The stereocilia of the different rows are joined by horizontal links just below their tips (arrowheads). The rows are also joined by tip links (arrows) that attach to the tip of the smaller stereocilia and to the upper density (double arrow-heads) of the taller stereocilia. (Scale bar is 200 nm). The insert shows a higher magnification of the lower tip link. (Scale bar is 100 nm). Adapted from Pickles (1988).

Figure 2.6 Detailed view of an inner hair cell(left) and an outer hair cell (right) with their supporting structures and nerve endings. Adapted from Pickles (1988).
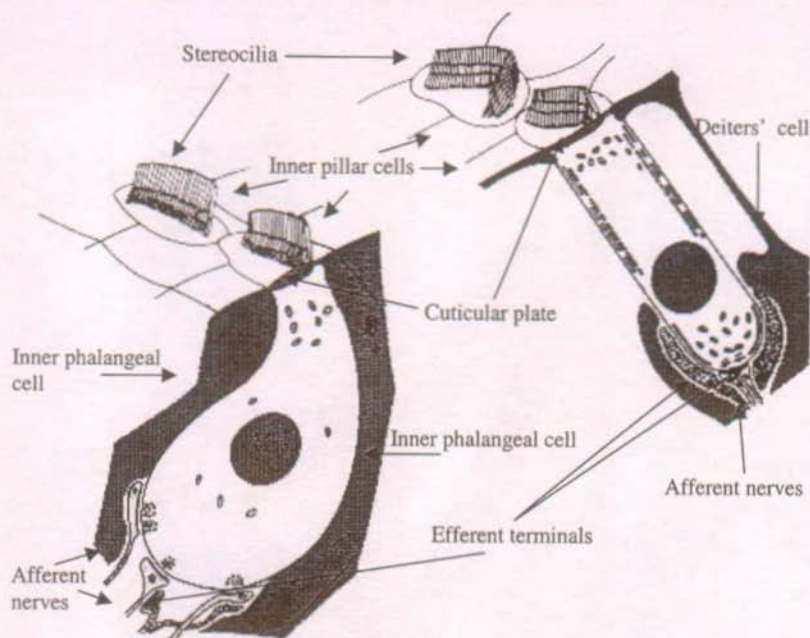
The innervation of the hair cells is illustrated in Figure 2.7. The majority of fibres in the auditory nerve contact the inner hair cells and are afferent fibres, i.e., fibres that carry the information to the brain. The afferent fibres from the inner hair cells at a certain point along the cochlea are bundled with the afferent fibres that contact a range of outer hair cells about 0.6 mm basalward. Most descending (efferent) fibres contact outer hair cells and come from the contralateral side of the brain. Synapses from both types of hair cells to the afferent fibres are excitatory, and synapses from the efferent projections to the outer hair cells are inhibitory. Very few efferent fibres innervate inner hair cells (not shown in Figure 2.7), but when they do they form an inhibitory axo-dendritic synapse on the dendrites of the spiral ganglion cell close to the inner hair cell synapses.
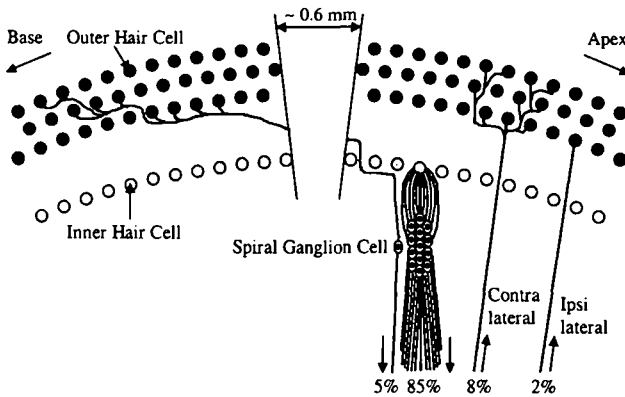
Figure 2.7 Innervation of the hair cells in humans. The percentages indicate the representation of fibres of the given type in the auditory nerve. Adapted from Watts (1993).

2.4 Function of the Cochlea

Three principal elements in the function of the cochlea may be distinguished, namely: the sound-induced motion of the basilar membrane in the fluid-filled cochlea, the transduction of this motion by the inner hair cells into a neural signal, and the neuro-mechanical feedback introduced by the outer hair cells. Since these elements are essential for the creation of a good model of the auditory periphery, we will discuss them in some detail.

2.4.1 Basilar membrane motion

Sound is input to the cochlea through motion of the oval window, created by movement of the stapes. Sinusoidal movement of the stapes causes a wave, which is propagated by the combined movement of the fluid and the basilar membrane. A three-dimensional representation of the basilar membrane displacement in the uncoiled cochlea is shown in Figure 2.8. The coiling of the cochlea has no significant effect on the travelling wave; its primary purpose appears to be to save space.
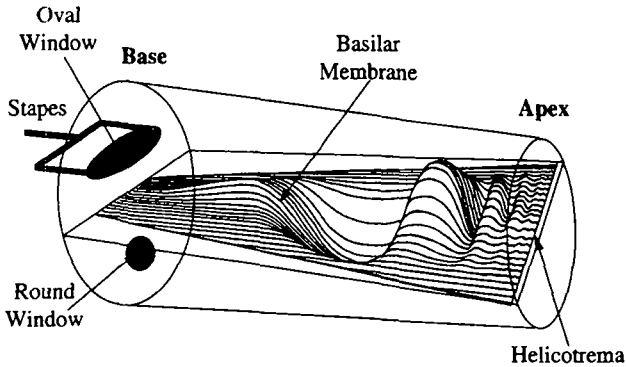
Figure 2.8 Three-dimensional representation of the travelling wave on the basilar membrane in the uncoiled cochlea. The amplitude of the displacement is largely exaggerated in this figure to show the form of the wave more clearly.

At the basal end of the cochlea, the basilar membrane is stiff, and the membrane-displacement wave propagates fast. As the wave travels down the cochlear canal, the stiffness of the basilar membrane decreases. This causes the wave to slow down, and its amplitude increases. Further down the cochlear canal, the membrane becomes too flexible to support a wave at the given frequency, and the energy of the wave is quickly dissipated. For each frequency there exists therefore a point on the basilar membrane for which the displacement of the basilar membrane is maximal. This point is called the "best place", and its position, measured from the oval window, increases logarithmically with decreasing frequency down to about 1 kHz. At lower frequencies the position varies more linearly with frequency. Conversely, for every point on the basilar membrane, there exists a frequency, the "best frequency", that induces maximum displacement of the basilar membrane at that point.
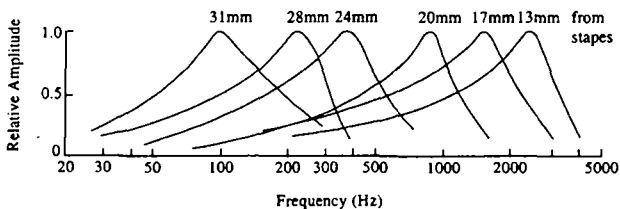

Figure 2.9 Frequency responses at six different points on the cochlear partition. Adapted from von Békésy (1960).

In Figure 2.9 the amplitude of the travelling wave envelope was measured as the stimulus frequency was varied with constant peak stapes displacement. If the data is recalculated for constant sound pressure level, i.e. constant peak stapes velocity, the low-frequency slopes become flat (Eldredge 1974). The cochlea thus appears to act purely as a poorly selective low-pass filter whose cut-off frequency decreases as one travels down the cochlea. Von Békésy's measurements, however, were performed on cadavers at very high sound intensity (130dB SPL). It is now known that not only does the animal have to be alive, but the cochlea must be in extremely good physiological condition to show a satisfactory mechanical response. More recent measurements by Johnstone, et al. (1986), using the Mössbauer technique, or Ruggero (1992), using Doppler-shift laser velocimetry, show a more complicated picture (see Figure 2.10).

In Figure 2.10 it may be seen that at low sound intensity the basilar membrane acts as a selective band-pass filter with high gain, whereas gain and selectivity decrease with increasing sound intensity. It may also be seen that well below the best frequency — e.g. below 6 kHz for the point shown in Figure 2.10 — the gain does not depend on stimulus intensity; the basilar membrane response is completely linear in this frequency range. It has been shown by Sellick et al. (1982) that the response of the basilar membrane after death of the animal is linear for all frequencies of stimulation and that the gain relative to the stapes velocity is substantially reduced around the best frequency of a given point along the basilar membrane. This seems to indicate that there is an active process in the cochlea which creates the high gain and selectivity. The most widely accepted hypothesis is that the outer hair cells, together with the organ of Corti and the basilar membrane, create an active mechanical amplifier. This hypothesis will be more fully discussed in paragraph 2.4.3.
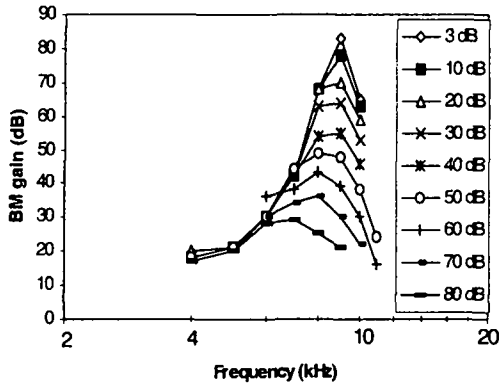
Figure 2.10 Frequency response at one point of the basilar membrane of a chinchilla cochlea for 9 different sound pressure levels. 0 dB SPL corresponds to an air pressure of 20 µPa at the eardrum. The gain is measured as the peak basilar membrane velocity divided by the peak stapes velocity, the later being constant for a given sound pressure level. Adapted from Ruggero (1992).

2.4.2 *Transduction by the inner hair cells.*

Since 95% of the outgoing connections from the cochlea come from the inner hair cells, we must presume that it is their task to convey information concerning the basilar membrane movement to the central nervous system. When the basilar membrane moves, the organ of Corti is displaced. The stereocilia of the outer hair cells, which are attached to the tectorial membrane, are displaced due to the shear between the reticular lamina (the rigid upper-surface of the organ of Corti) and the tectorial membrane. The stereocilia of the inner hair cells do not touch the tectorial membrane, but fit loosely into a raised groove known as Hensen's stripe on the under-surface of the tectorial membrane. When the reticular lamina moves with the basilar membrane forces are exerted on the stereocilia, mainly due to the viscous drag of the endolymph. The displacement of the inner hair cell stereocilia is thus proportional to the velocity of the basilar membrane motion.
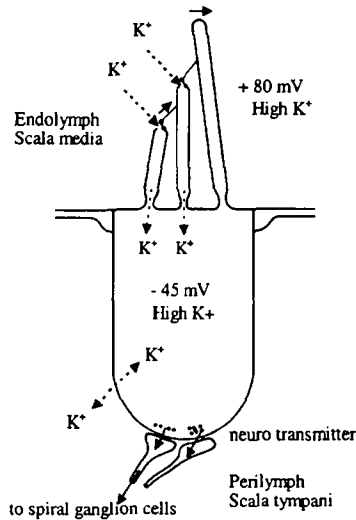
Figure 2.11 Current opinion on the transduction by the inner hair cells assumes that the tip links of a stereocilia control the probability that the ion channels on this stereocilia are open. There are only a few ion channels (1-4) per stereocilia, and they are probably located at the tip of the stereocilia. Adapted from Pickles (1988).

In the light of recent evidence (for instance Howard, et al. 1988), the mechano-electrical transduction by the hair cells can be dealt with in terms of variable resistances and a modern version of the Davis battery theory (Davis 1958). In this theory, the ion channels in the tips of the stereocilia act as variable resistances. Ions flow into the cell, driven by the battery of the endolymphatic potential and the intracellular potential. Intracellular depolarisation causes release of transmitter, and activation of the auditory nerve fibres via the spiral ganglion cells.

Potassium is most abundant both intracellularly and extracellularly at the hairs (see Figure 2.11). Furthermore, the ion channels of the stereocilia are permeable by potassium. The ion flows will therefore be dominated by the potassium current. Because the basal ends of the hair cells are in contact with perilymph, which has a low potassium concentration, potassium entering the hair cell at the apex will automatically diffuse out of the cell at the basal end. The hair cells thus modulate the potassium current flow from the scala media to the scala tympani.

There seem to be only a few ion channels per stereocilia; these oscillate spontaneously between their closed and open state, probably under influence of thermal energy. At rest, the channels are in their open state about 20% of the time. Displacement of the stereocilia in the direction of the tallest stereocilia

stretches the tip-links and thereby increases the amount of time they are open. Displacement in the opposite direction naturally has the opposite effect. Since the channels cannot be more than 100% open or 100% closed, the cell's response, i.e., the intracellular potential, will vary between two asymptotes. As can be seen in Figure 2.12, the variation between these two asymptotes has a form which resembles a sigmoid function.

The fact that the inner hair cell reacts more strongly in one direction than the other creates, for a periodic stimulation, an average response which is larger than the resting voltage by an amount which depends on the amplitude of the input signal as long as the input signal does not drive the cell into saturation. The membrane of the inner hair cell also has a certain capacitance, and the leakage through the membrane introduces a resistance in parallel with this capacitance. Furthermore, the opening and closing of the channels in the stereocilia modulates the resistance between the inside of the inner hair cell and the positive voltage delivered by the endolymph. The inner hair cell therefore acts as a low-pass filter. This means that the amplitude of the variation around the average voltage will decrease with increasing frequency for a given stimulus intensity. We can see this effect clearly in Figure 2.13, which shows measurements of the intracellular voltage of the inner hair cell for different stimulation frequencies.
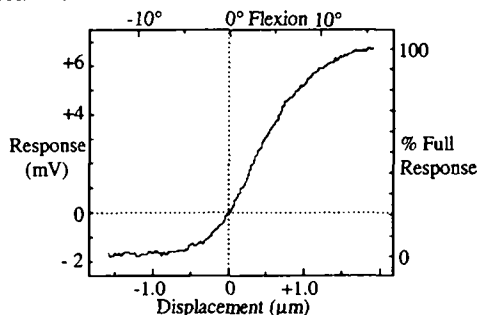
Figure 2.12 Hair cell response as function of hair bundle deflexion. Adapted from Hudspeth and Corey (1977).
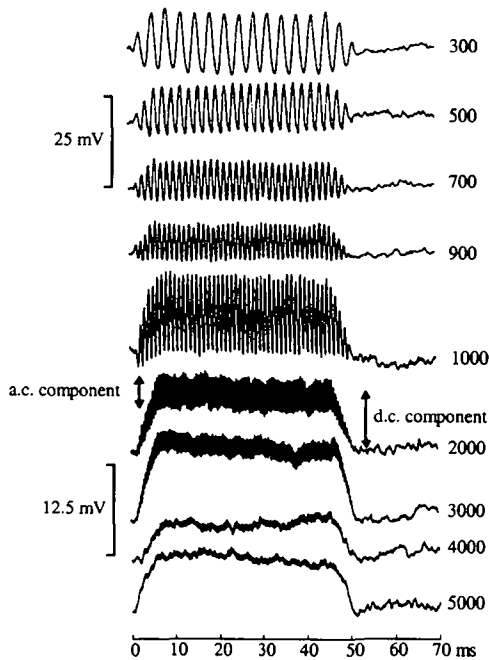
Figure 2.13 Intracellular voltage changes in an inner hair cell for different frequencies of stimulation. Note the change of scale for the lower five traces. Adapted from Palmer and Russell (1986).
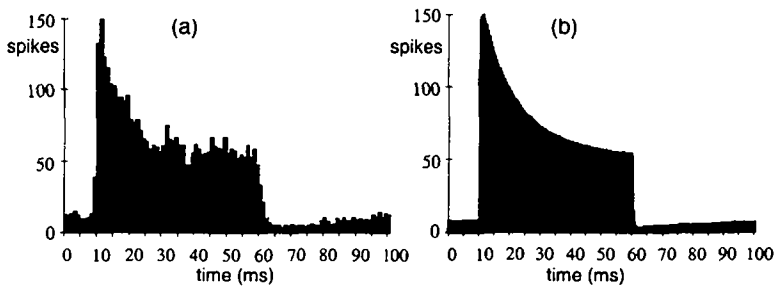
Figure 2.14 Post stimulus time histogram of a single auditory nerve fibre in response to repeated stimulations with a pure tone as measured in the cat (Pickles 1988), and as predicted by Meddis' hair cell model (Meddis 1986). These histograms keep the temporal relation between the occurrence of the spikes and the stimulus tone.

In humans each inner hair cell synapses with about 20 spiral ganglion cells, whose axons form the auditory nerve. The spiral ganglion cells transmit the

information transduced by the inner hair cells to the brain by means of voltage pulses (spikes). As we will discuss in chapter 4, most neurons interact through chemical synapses by releasing a special chemical substance (neurotransmitter) into the synaptic cleft, which is the space between the two cells. Each time an inner hair cell releases a packet of neurotransmitter into the synaptic cleft, the spiral ganglion cell on the other side of the cleft spikes. The release of neurotransmitter by the inner hair cell is not a deterministic process, but the probability of release is proportional to the intracellular voltage change and the amount of neurotransmitter available. When the inner hair cell is stimulated, the neurotransmitter is used partially up, so that the response to the next stimulus is reduced. This results in a decreasing response of the spiral ganglion cells to a continuous stimulation of the inner hair cell. A good model of the transmitter release by the inner hair cell has been developed by Meddis (Meddis, 1986, 1988; Lopez-Poveda, et al., 1997). Typical post stimulus time histograms of responses of a cat auditory nerve fibre, and of a model fibre, to a high frequency stimulus are given in Figure 2.14. Figure 2.14a clearly shows the adaptation of the auditory nerve fibres response to the stimulus. Adaptation effects of different origins than the consumption of neurotransmitter have also been shown in hair cells. However, since an adaptation effect is hardly visible in the intracellular voltage changes shown in Figure 2.13, the adaptation seen on the auditory nerve signal must be dominated by the conversion of the intracellular voltage into a neural signal, i.e., by the neurotransmitter release.

2.4.3 Function of the outer hair cells.

Although there are about 3.5 times more outer hair cells than inner hair cells, only 5% of the afferent nerve fibres transmit information from the outer hair cells to the brain. Furthermore, the information on these fibres is of very low spatial and temporal resolution. It must therefore be concluded that the function of the outer hair cells is not to signal the basilar membrane movement to the central nervous system. The few fibres that project from the outer hair cells to the brain probably only carry information about the operating point of the outer hair cells. It has been hypothesised that the outer hair cells act as an active mechanical amplifier, of which the gain is controlled by feedback connections from the central nervous system. The actual details of the way the outer hair cells perform this task are not yet understood, and there is only indirect support for this hypothesis.

The strongest indication, initially shown by Kiang et al. (1970) in the cat, is that selective damage to the outer hair cells reduces the selectivity and gain of the neural tuning curves around the best frequency, but leaves the low frequency part unaffected. The same effect may be obtained by stimulation of the crossed olivocochlear bundle, which descends from the contralateral

superior olivary complex to inhibit mainly the outer hair cells. Both experiments are not conclusive though, because it is unclear whether only the outer hair cells are affected in these experiments, and it has not been shown that the gain of the basilar membrane movement has been reduced, but only that the response of the auditory nerve fibres has been influenced.

The aforementioned experiments indicate that the outer hair cells function in some way necessary for the sharp tuning and sensitivity of the inner hair cells. They do not prove that this process is necessarily active. The strongest direct evidence for the existence of active mechanical processes in the cochlea comes from the observation that under certain circumstances the cochlea can actually generate sound (Pickles, 1988). Furthermore, there is evidence that this generation of sound is related to the activity of the hair cells themselves.

Finally, various forms of motility have been demonstrated in isolated outer hair cells. It has been shown that outer hair cells can change their length when electrically stimulated (see for instance Ashmore, 1987), and the anatomy of the organ of Corti, in which the outer hair cells are attached only at both ends, does not impede this movement. It has also been shown, by Crawford et al. (1985), that the bundle of stereocilia of a hair cell from a turtle cochlea was able to oscillate in synchrony with an electrical oscillation inside the cell produced by a stimulation of the stereocilia at low frequency. It is clear that with the slower forms of motility the outer hair cells can control the mechanical operating point of the organ of Corti. It is still unclear whether the faster forms of motility are fast enough to be able to act on a cycle-by-cycle basis, which is needed to produce the high gain observed at low stimulus intensity.

2.5 Summary

The human ear, which is the first stage in our auditory system, translates the mechanical vibrations of sound, i.e., a time-varying pressure signal at the eardrum, into a time-varying pattern of excitation on the auditory nerve. The ear can be divided into three parts.

The outer ear captures the incoming sound and the vibration of the eardrum is transferred by the bones in the middle ear to the oval window of the inner ear.

The most important structure in the ear, from a modelling point of view, is the cochlea, the fluid-filled, coiled structure of the inner ear. The cochlea is divided in two by the basilar membrane. Displacement of the oval window by the middle ear bones creates a pressure wave in the fluids of the cochlea and the cochlea filters this wave, so that different sound frequencies will cause different parts of the basilar membrane to vibrate.

The base of the basilar membrane is most sensitive to high frequencies and the apex to low frequencies. In between, the best frequency decreases approximately exponentially with position along the cochlea.

The inner hair cells sense the vibration of the basilar membrane all along its length and transduce this motion into electrical signals. These electrical signals are then chemically transmitted to the spiral ganglion cells, whose axons are the auditory nerve fibres that carry the information from the ear to the brain.

The outer hair cells locally control the gain of the cochlear filter under control of descending connections from the brain.

2.6 References for chapter 2

Ashmore, J.F. (1987) "A fast motile response in guinea-pig outer hair cells: the cellular basis of the cochlear amplifier," *Journal of Physiology*, Vol. 388, pp. 323-347.

Békésy, G. von. (1960) *Experiments in Hearing*. McGraw-Hill.

Brownell, W.E. and Shehata, W.E. (1990) "The effect of cytoplasmic turgor pressure on the static and dynamic mechanical properties of the outer hair cells," in P. Dallos, C.D. Geisler, J.W. Matthews, M.A. Ruggero, and C.R. Steele, editors, *The Mechanics and Biophysics of Hearing*, Springer-Verlag, Berlin, pp. 52-60.

Crawford, A.C. and Fettiplace, R. (1985) "The mechanical properties of ciliary bundles of turtle cochlear hair cells," *Journal of Physiology*, Vol. 364, pp. 359-379.

Dallos, P. (1984) "Peripheral mechanisms of hearing," *American Physiological Society*, Bethesda, 1984, pp. 595-637.

Davis, H. (1958) "Transmission and transduction in the cochlea," *Laryngoscope*, Vol. 68, pp. 359-382.

Eldredge, D.H. (1974) "Inner ear - cochlear mechanics and cochlear potentials," in W.D. Keidel and W.D. Neff, editors, *Handbook of Sensory Physiology*, Springer-Verlag, Berlin, Vol. 5/1, pp. 549-584.

Evans, E.F. (1982) "Functional anatomy of the auditory system," in H.B. Barlow and J.D. Mollon, eds, *The Senses*, Cambridge University Press, pp. 251-306.

Howard, J., Roberts, W.M., and Hudspeth, A.J. (1988) "Mechanoelectrical transduction by hair cells," *Annual Reviews in Biophysics and Biophysical Chemistry*, Vol. 17, pp. 99-124.

Hudspeth, A.J. and Corey, D.P. (1977) "Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli," *Proceedings of the National Academy of Sciences of the USA*, Vol. 74, pp. 2407-2411.

Johnstone, B.M., Patuzzi, R. and Yates, G.K. (1986) "Basilar membrane measurements and the travelling wave," *Hearing Research*, Vol. 22, pp. 147-153.

Kelly, J.P. (1991) "Hearing." In E.R. Kandel, J.H. Schwartz, and T.M. Jessell, editors, *Principles of Neural Science*, Prentice Hall International, pp. 481-499.

Kessel, R.G. and Kardon, R.H. (1979) *Tissues and Organs: A Text-Atlas of Scanning Electron Microscopy*, W.H. Freeman, New York.

Kiang, N.Y.-S., Moxon, E.C., and Levine, R.A. (1970) "Auditory-nerve activity in cats with normal and abnormal cochleas," In G.E.W. Wolstenholme and J. Knight, editors, *Sensorineural Hearing Loss*, CIBA Foundation Symposium. Churchill, London, pp. 241-268.

Lopez-Poveda, E.A., O'Mard, L.P., and Meddis, R. (1997) "A revised computational inner hair cell model," *Proceedings of the 11th International Symposium on Hearing*, Whurr Publishers, London, pp. 102-108.

Meddis, R. (1986) "Simulation of mechanical to neural transduction in the auditory receptor," *Journal of the Acoustical Society of America*, Vol. 79, pp. 702-710.

Meddis, R. (1988) "Simulation of auditory-neural transduction: Further studies," *Journal of the Acoustical Society of America*, Vol. 83, pp. 1056-1062.

Nedzelnitsky, V. (1980) "Sound pressures in the basal turn of the cat cochlea," *Journal of the Acoustical Society of America*, Vol. 68, pp. 1676-1689.

Palmer, A.R., and Russell, I.J. (1986) "Phase-locking in the cochlear nerve of the Guinea-pig and its relation to the receptor potential of inner hair cells," *Hearing Research*, Vol. 24, pp. 1-15.

Pickles, J.O., (1988) *An Introduction to the Physiology of Hearing*, Academic Press, London.

Ruggero, M.A. (1992) "Responses to sound of the basilar membrane of the mammelian cochlea," *Current Opinion in Neurobiology*, Vol. 2, pp. 449-456.

Sellick, P.M., Patuzzi, R., and Johnstone, B.M. (1982) "Measurement of basilar membrane motion in the guinea pig using the Mössbauer technique," *Journal of the Acoustical Society of America*, Vol. 72, pp. 131-141.

Watts, L. (1993) *Cochlear Mechanics: Analysis and Analog VLSI*, Ph.D. thesis, California Institute of Technology, Pasadena.

3. The Electronic Ear

3.1 Introduction

In the previous chapter we have identified the cochlea as the most important sound processing structure in the ear. It filters sound into separate spectral components and generates auditory nerve signals which transmit acoustical information to the brain. In this chapter we will concentrate on modelling the key elements of the cochlear function, i.e., the mechanical filtering by the basilar membrane and the transduction of this motion by the inner hair cells.

3.2 The basilar membrane model

3.2.1 Background

Once sound is transferred by the middle ear from the eardrum to the cochlea, it initiates travelling waves of fluid pressure and basilar membrane motion which propagate from base to apex. For a given frequency, the wave travels almost loss-less through the basal part of the cochlea, where the basilar membrane is relatively stiff. As the stiffness of the basilar membrane decreases the wave slows down and energy starts to accumulate, yielding an increase in gain at the 'best place' for this frequency. More apical of this 'best place' the membrane becomes too elastic for a wave of this frequency and the wave dies out quickly. The simplest way to capture this behaviour in an electronic implementation is with a cascade of scaled second-order low-pass filters. A wave travels loss-less at the start of the cascade, has a certain gain at the 'best place', and dies out quickly after the best place. The first "analog electronic cochlea", published by Lyon and Mead in 1988, was based on this idea. Since then several other analogue VLSI models have been proposed which try to capture more of the details of the biological cochlear function (Lyon, 1991; Liu et al., 1991; Sarpeshkar et al., 1996; Watts et al., 1992; Watts, 1993). The design of the basilar membrane model in this section is based on an improved version of Lyon's original analogue electronic cochlea, as presented by Watts et al. (1992). I have chosen this realisation, because it is the simplest and thus smallest cochlear model, allowing for a large number of filters in the cascade,

and because a certain amount of experience has accumulated over the years with this implementation, published by Watts et al. (1992).

In the following sections we shall first discuss the second-order low-pass filter which is in the filter cascade and is identical to the filter used by Watts et al. The improvements I have made to the silicon cochlea will then be presented. These are a new implementation of the differentiation of the low-pass filter output, and a new implementation of the exponential bias current generation for the silicon cochlea. Finally, measurements of the silicon cochlea will be presented and discussed.

3.2.2 The transconductance amplifier

A standard building block that will be used for the implementation of the second-order low-pass filter is the transconductance amplifier. Figure 3.1a shows the circuit of such a transconductance amplifier. When biased in weak inversion it has a hyperbolic tangent transfer function given by :

$$I_{out} = I_{bias} \tanh\left(\frac{V_+ - V_-}{2 \, n \, U_T}\right) \tag{1}$$

with I_{bias}, V_+ and V_- as in Figure 3.1, and n is a slope factor between 1 and 2. The thermal voltage $U_T = kT/q$, with k the Boltzmann constant, T the temperature in Kelvin, and q the charge of an electron. U_T is about 25 mV at room temperature.

The transconductance amplifier is biased in weak inversion when the current through the differential pair transistors T_1 and T_2 is much smaller than their specific current, i.e.:

$$I_{bias} << I_S = 2 \, n \, \mu \, C_{ox} \, W/L \, U_T^2 \tag{2}$$

where W is the channel width of the transistor, L the length, μ the mobility of the minority carriers, and C_{ox} the gate oxide capacitance per unit area.

For small inputs ($|V_+ - V_-| < 60$ mV) we can approximate the amplifier as a linear transconductance:

$$I_{out} = g_a (V_+ - V_-) \tag{3}$$

with the transconductance given by:

$$g_a = \frac{I_{bias}}{2 \, n \, U_T} \tag{4}$$

This last equation shows that the transconductance of the amplifier is proportional to the bias current when the amplifier operates in weak inversion.

For auditory processing, the amplifiers will necessarily operate in the weak inversion range. When we want to build filters for frequencies in the auditory range, we will need long time constants, thus either large capacitors or small conductances. Since it is not reasonable to create large capacitors on-chip, we will have to create small conductances, i.e., use small bias currents. These small currents will thus bias the amplifier in the weak inversion range, unless we make the W/L ratio of the transistors very small. That, however, would mean using extremely long transistors though, which would lead to a large area consumption and is therefore not a practical option.
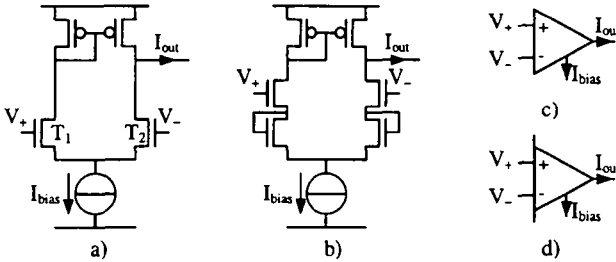

Figure 3.1 A basic transconductance amplifier: a) schematic and c) symbol; a transconductance amplifier with an enlarged linear range: b) schematic and d) symbol.

Watts, et al., (1992) also used a transconductance amplifier with an enlarged linear range as shown in Figure 3.1b. The transfer function of this amplifier is given by:

$$I_{out} = I_{bias} \tanh\left(\frac{V_+ - V_-}{2\,n\,(n+1)\,U_T}\right) \tag{5}$$

and

$$g_a = \frac{I_{bias}}{2\,n\,(n+1)\,U_T} \tag{6}$$

The linear range of this amplifier is thus enlarged by a factor $n + 1$, where n is about 1.5, and the transconductance is reduced by a factor $n + 1$ with respect to the normal transconductance amplifier with the same bias current.

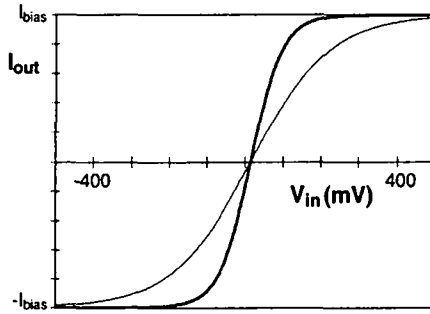The transfer functions of both amplifiers are shown in Figure 3.2.

Figure 3.2 Transfer function of the transconductance amplifier (heavy line) and the modified transconductance amplifier.

3.2.3 The second-order low-pass filter

A second-order low-pass filter is made of three transconductance amplifiers (A_1, A_2, A_3 as shown in Figure 3.3) and two capacitors.
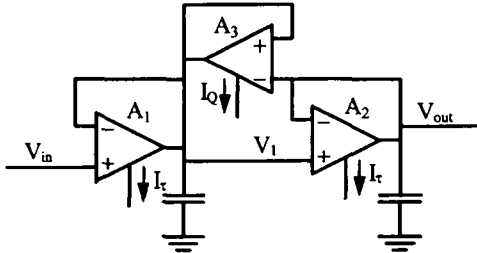
Figure 3.3 One section of the cochlear cascade, with differentiator.

The transfer function of this filter is given in the Laplace domain by:

$$H(s) = \frac{V_{out}}{V_{in}} = \frac{1}{1 + \tau s/Q + (\tau s)^2} \tag{7}$$

where $s = j\,\omega$, $j^2 = -1$, and ω is the angular frequency, the time constant $\tau = C/g_\tau$ when both A_1 and A_2 have a conductance g_τ and both capacitors have capacitance C. The quality factor Q of the filter can be expressed as:

$$Q = \frac{1}{2 - g_Q/g_\tau} \tag{8}$$

where g_Q is the conductance of the amplifier A_3.

The gain and phase response of the filter described by equation (7) are shown in Figure 3.4 for two values of Q.

Figure 3.4 a) Gain and b) phase response of the second-order low-pass filter for Q=1 (heavy line) and Q=10 when $1/\tau = 1000$ s^{-1}.

3.2.4 Stability of the filter

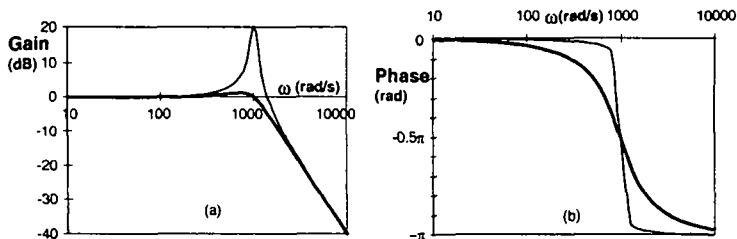From equation (8) we can see that when $g_Q = 2\ g_\tau$, Q becomes infinite. From equation (7) we can then see that when $\omega = 1/\tau$, the gain of the filter becomes infinite and the filter will be unstable. This gives $g_Q < 2\ g_\tau$ as the small-signal stability limit of the filter.

The filter of Figure 3.3 also has a large-signal stability limit. In order to obtain the transfer function of equation (7) we have treated the filter as a linear system. This approximation is only valid for small input signals. It has been shown by Mead (1989) that large transient input signals can create a sustained oscillation in this filter. During most of this oscillation all three amplifiers are saturated so that their output is either plus or minus their bias current. We can therefore adopt a piece-wise linear approach in which we treat the amplifiers as current sources to analyse this behaviour. The following analysis is mostly adapted from Mead's, but is more general since it does not assume that the amplitude of the oscillation is equal to the supply voltage.

When V_{in} suddenly increases by a large amount, the amplifier A_1 will saturate and will charge the capacitor at its output with the maximum output current I_τ. If, at the same time, V_1 is larger than V_{out}, A_3 will also charge the capacitor with its maximum output current I_Q and we can write for V_1:

$$\frac{dV_1}{dt} = \frac{I_Q + I_\tau}{C} \qquad \text{where } V_{out} \ll V_1 \ll V_{in} \qquad (9)$$

V_1 will thus rise at its maximum rate. Once V_1 catches up with V_{in}, the output current of A_1 changes sign and we write for V_1:

$$\frac{dV_1}{dt} = \frac{I_Q - I_\tau}{C} \qquad \text{where } V_{out} \ll V_{in} \ll V_1 \qquad (10)$$

In order to have Q larger than one, I_Q has to be larger than I_τ so that in this case V_1 will continue to increase with a smaller slope until it reaches the positive power supply V_{DD} or until V_{out} catches up with V_1. As long as V_1 stays larger than V_{out} we can write in our piece-wise linear approach for V_{out}:

$$\frac{dV_{out}}{dt} = \frac{I_\tau}{C} \qquad \text{where } V_{out} \ll V_1 \qquad (11)$$

Once V_{out} catches up with V_1, the sign of the output of A_3 will change, and V_1 will start its steep descent, until V_1 goes below V_{in}, when the sign of the output of A_1 changes and V_1 descents more slowly to the negative power supply V_{SS}, or until V_{out} catches up again. Figure 3.5 sketches the behaviour of the circuit according to the above equations. The thick line shows the evolution of V_1 and the thin line shows the same for V_{out}. Whenever V_{out} catches up with V_1, the change in both voltages will change direction.
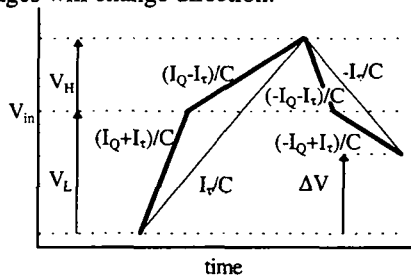
Figure 3.5 Piece-wise linear approximation of the waveform for V_1 (bold) and V_{out}. The *slope of each line is indicated next to it.*

By comparing the voltages at which V_{out} catches up with V_1 at the start and the end of a single rise and fall cycle, we can determine the nature of the oscillation. If ΔV (see Figure 3.5) is positive, the amplitude of the oscillation will decrease during each period and the oscillation will cease after a certain time. The limit of stability is reached when ΔV becomes zero, so that the amplitude of the oscillation stays constant. For the rising part of the oscillation we write:

$$\frac{C}{I_Q + I_\tau} V_L + \frac{C}{I_Q - I_\tau} V_H = \frac{C}{I_\tau}(V_L + V_H) \qquad (12)$$

where V_L and V_H are as shown in Figure 3.5. Similarly, for the falling part, we write when ΔV equals zero:

$$\frac{C}{I_Q + I_\tau} V_H + \frac{C}{I_Q - I_\tau} V_L = \frac{C}{I_\tau} (V_L + V_H) \tag{13}$$

Equations (12) and (13) can only be satisfied when $V_L = V_H$. Substituting V for V_L and V_H in either equation, and dividing by CV/I_τ yields:

$$\frac{1}{I_Q/I_\tau + 1} + \frac{1}{I_Q/I_\tau - 1} = 2 \tag{14}$$

Rewriting this equation we obtain the following solution:

$$\frac{I_Q^2}{I_\tau^2} - \frac{I_Q}{I_\tau} - 1 = 0 \ => \ \frac{I_Q}{I_\tau} = \frac{1 + \sqrt{5}}{2} \approx 1.62 \tag{15}$$

This gives the critical value for large signal stability of the low-pass filter of Figure 3.3. Since the conductance of the amplifiers is directly proportional to the bias currents, this large-signal stability condition also limits g_Q/g_τ to this value, and thus limits Q to a maximum value of 2.63 (see equation (8)). The large-signal stability limit therefore severely limits the maximum quality factor of the filter.

3.2.5 The stabilised second-order low-pass filter

The filter can be improved by using two wide-range transconductance amplifiers to implement A_1 and A_2 (Figure 3.6) and a basic transconductance amplifier for A_3 (Watts, et al, 1992). In this case we can write for the conductance ratio:

$$\frac{g_Q}{g_\tau} = (n + 1) \frac{I_Q}{I_\tau} \tag{16}$$

which ensures that g_Q/g_τ becomes 2 before I_Q/I_τ becomes 1.62, since n is larger than 1, so that the filter is always large-signal stable whenever it is small-signal stable.
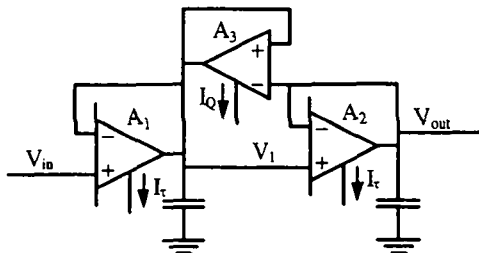

Figure 3.6 Modified second-order low-pass filter.

With the two wide-range transconductance amplifiers A_1 and A_2 and the one normal transconductance amplifier A_3 we have a second-order low-pass filter for which we can set the cut-off frequency and the quality factor using the bias currents of these amplifiers. By cascading these filters, and biasing the amplifiers with exponentially decreasing currents we can create a model of the basilar membrane.

3.2.6 Differentiation

The voltage V_{out} (Figure 3.6) at the output of each second-order stage in the cochlear filter cascade represents the displacement of a small section of the basilar membrane. However, since the stimulation of the inner hair cells in the biological cochlea is proportional to the velocity of the basilar membrane, the output of each second-order stage has to be differentiated. This can be done by creating a copy of the output current I_{dif} of amplifier A_2 at every stage as in Watts, et al. (1992). Since the voltage on a capacitor is proportional to the integral of the current onto the capacitor, I_{dif} is effectively proportional to the basilar membrane velocity. Yet, with equal displacement amplitudes, velocity will be much larger for high frequencies than for low frequencies, yielding output signals with amplitudes that decrease from the beginning of the cochlea to the end. This can be corrected by normalising I_{dif} to give equal amplitude at every output. A second resistive line with identical tilt controlling the gain of the current mirrors that create the copies of I_{dif} at each stage is used for this purpose by Watts, et al. However, relying on an identical tilt on the second resistive line introduces an extra source of mismatch in the circuit.
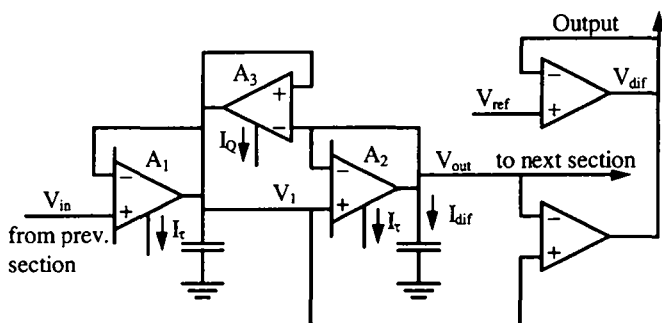

Figure 3.7 One section of the cochlear cascade, with differentiator.

An alternative solution which does not need normalisation is to take the difference between V_{out} and V_1 (see Figure 3.7). We can rewrite equation (3) applied to A_2 as:

$$g_\tau (V_1 - V_{out}) = I_{dif} \tag{3}$$

or:

$$V_1 - V_{out} = \frac{I_{dif}}{g_\tau} = \frac{s\,C\,V_{out}}{g_\tau} = \tau\,s\,V_{out} \tag{17}$$

This is equivalent to differentiating V_{out}, with 0 dB gain at the cut-off frequency for all stages. Figure 3.8 shows the gain and phase response of the filter after differentiation. We can see that a single filter only has a shallow high frequency cut-off slope of 20 dB per decade. In the filter cascade, however, the 40 dB per decade cut-off slopes (Figure 3.4) of the individual low-pass filters will be accumulated. This can yield very steep high frequency cut-off slopes, as we will see in the measurements later on in Figure 3.12.
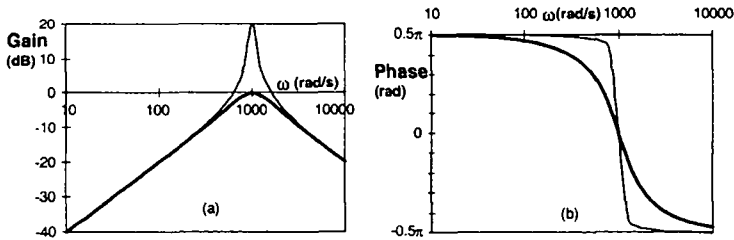

Figure 3.8 a) Gain and b) phase response of the second-order low-pass filter with differentiator for Q=1 (heavy line) and Q=10 when $1/\tau = 1000\ s^{-1}$.

The subtraction of equation (17) can be implemented with a combination of 2 transconductance amplifiers, as shown in Figure 3.7, which ensure that:

$$V_{dif} - V_{ref} = V_1 - V_{out} = \tau\,s\,V_{out} \tag{18}$$

Since V_{ref} is constant, V_{dif} will vary around V_{ref} by an amount equal to τ s V_{out}.

The two amplifiers implementing this can have a large bias current, so they may also be used to buffer the cascade voltages before connecting them to the output pins of the chip, to avoid charging the cochlear cascade with the extra capacitance introduced by the output pins.

3.2.7 Exponential bias generation

Since there is an exponential relationship between position along the basilar membrane and best frequency in the real cochlea, we will need to use filters with exponentially decreasing cut-off frequencies in our model. In all the silicon cochlear models mentioned in the introduction of this chapter, the exponential dependency is obtained using a linear decreasing voltage on the gates of MOS transistors operating in weak-inversion. In weak-inversion, the drain current of

a saturated nMOS transistor with its source tied to the bulk, and its gate voltage referred to the same bulk, can be expressed by:

$$I_D = I_S \, e^{\frac{V_G - V_{T0}}{n U_T}} \tag{19}$$

with I_S as defined in equation (2) and V_{T0} the threshold voltage of the transistor. This shows that the drain current depends exponentially on the gate voltage. A spatial voltage distribution which decreases linearly with distance is easily created using a resistive polysilicon line; if there is a voltage difference between the two ends of the line, the voltage on the line will decrease linearly all along its length. It is therefore possible to create a filter cascade with an exponentially decreasing cut-off frequency by biasing the amplifiers of Figure 3.6 using MOS transistors whose gates are connected by equal lengths of the polysilicon line (Lyon and Mead, 1988).

As we can see in equation (19), however, the drain current also depends exponentially on the threshold voltage and small variations in V_{T0} will introduce large variations in the drain current. Because both the cut-off frequency and the quality factor of the filters are proportional to these drain currents, large parameter variations are generated by small V_{T0} variations. An RMS mismatch of 12% in the drain current of two identical transistors with equal gate and source voltages is not exceptional (Vittoz, 1985), even when sufficient precautions are taken.

I have circumvented this problem by using CMOS Compatible Lateral Bipolar Transistors (CLBTs) as bias transistors. A CLBT is obtained if the drain or source junction of a MOS transistor is forward-biased in order to inject minority carriers into the local substrate. If the gate voltage is negative enough (for an n-channel device), then no current can flow at the surface and the operation is purely bipolar (Vittoz, 1983; Arreguit, 1989). Figure 3.9 shows the major flows of current carriers in this mode of operation, with the source, drain and well terminals renamed emitter E, collector C and base B.
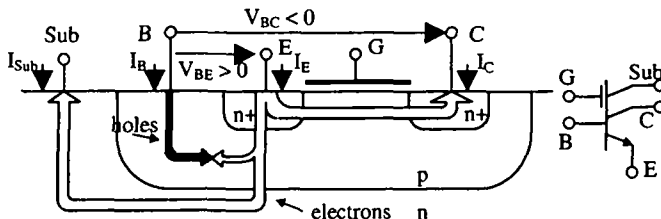

Figure 3.9 Bipolar operation of the MOS transistor : carrier flows and symbol.

Since there is no p+ buried layer to prevent injection to the substrate, this lateral npn bipolar transistor is combined with a vertical npn. The emitter

current I_E is thus split into a base current I_B, a lateral collector current I_C and a substrate collector current I_{Sub}. Therefore, the common-base current gain $\alpha = -I_C/I_E$ cannot be close to 1. However, due to the very small rate of recombination inside the well and to the high emitter efficiency, the common-emitter current gain $\beta = I_C/I_B$ can be large. Maximum values of α and β are obtained in concentric structures using a minimum size emitter surrounded by the collector and a minimum lateral base width.

For $V_{CE} = V_{BE}-V_{BC}$ larger than a few hundred millivolts, this transistor is in active mode and the collector current is given, as for a normal bipolar transistor, by

$$I_C = I_{Sb}\, e^{\frac{V_{BE}}{U_T}} \tag{20}$$

where I_{Sb} is the specific current in bipolar mode, proportional to the cross-section of the emitter-to-collector flow of carriers. Since I_C is independent of the MOS transistor threshold voltage V_{T0}, the main source of mismatch of distributed MOS current sources is suppressed, when CLBTs are used to create the current sources.

A disadvantage of the CLBT is its low Early voltage, i.e., the device has a low output resistance. Therefore, it is preferable to use a cascode circuit as shown in Figure 3.10. This yields an output resistance several hundred times larger than that of the single CLBT, whereas the area penalty, in a layout as shown in Figure 3.10b, is acceptable (Arreguit, 1989).
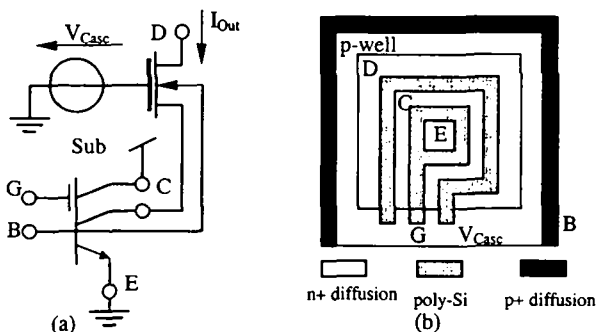

Figure 3.10 CLBT cascode circuit (a) and its layout (b).

Another disadvantage of CLBTs, when biased using a resistive line, is their base current, which introduces an additional voltage drop on the resistive line. However, since the cut-off frequencies in the cochlea are controlled by the output current of the CLBTs and since these cut-off frequencies are relatively

small (typically 20 kHz or less), the output current of the CLBTs will be small. If the common-emitter current gain β is much larger than 1, the base current of these CLBTs will be very small compared to the current flowing through the resistive line and the voltage error introduced by the small base currents will be negligible. Furthermore, since the cut-off frequencies of the cochlea will typically span 2 decades with an exponentially decreasing cut-off frequency from the beginning to the end, only the first few filters will have any noticeable influence on the current drawn from the resistive line.

3.2.8 The silicon cochlea

The final design of the silicon cochlea is shown in Figure 3.11. It uses the cochlear sections shown in Figure 3.7, CLBTs as the bias transistors of each filter, and one resistive line to bias all CLBTs.
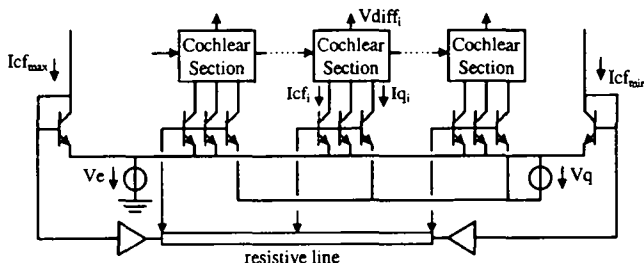

Figure 3.11 The improved silicon cochlea.

The cut-off frequencies of the first and the last low-pass filters in the cascade can be set by applying voltages to both ends of the resistive line, and the intermediate filters will have a cut-off frequency decreasing exponentially from the beginning to the end. Yet, if we directly apply a voltage to the ends of the resistive line, the actual cut-off frequency obtained will depend on the temperature, since the current depends exponentially on the applied voltage *normalised* to the thermal voltage U_T. It is therefore better to create the voltages at both ends of the resistive line on-chip using a current biasing a CLBT with its base connected to its collector. If this base voltage is buffered, so that the current through the resistive line is not drawn from the input current, the bias currents of the first and last filter — and thus the cut-off frequencies of all filters — can be set, independent of temperature.

A similar structure is used to create the voltage source V_q to control, independent of temperature, the actual quality factor of each section. The actual bipolar current mirrors implemented use the cascode structure shown in Figure 3.10a; however, this is not shown in Figure 3.11 for clarity.

3.2.9 Test results

The proposed silicon cochlea has been integrated using the ECPD15 (1.5 μm) technology of ES2 and contains 104 second-order stages on a 4.77mm × 3.21mm die. Every other stage is connected to a pin so its output voltage can be measured. In Figure 3.12, frequency response curves after on-chip derivation are shown for the output taps of both the cochlea described by Watts, et al. (1992) (a), and the improved version (b). This clearly shows the improved regularity of the cut-off frequencies and the gain obtained using CLBTs. The drop-off in gain for the higher frequency stages Figure 3.12(b) is a border effect, since at the beginning of the cochlea no accumulation of gain has yet taken place. In the figure on the left this is not visible, since the first nine outputs are not presented.
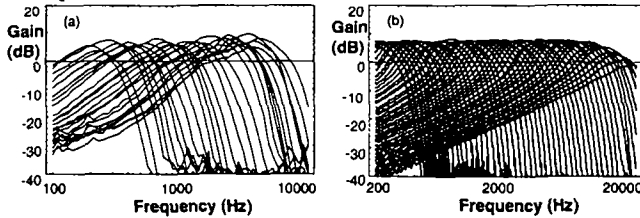
Figure 3.12 Measured frequency responses at the different taps for a) the original cochlea by Watts, et al. and b) the improved version.

Figure 3.13 shows the cut-off frequency versus tap number of both chips. Ideally, this should be a straight line on a log-linear scale, since the cut-off frequency decreases exponentially with tap number. This also clearly shows the improved regularity using CLBTs as current sources.
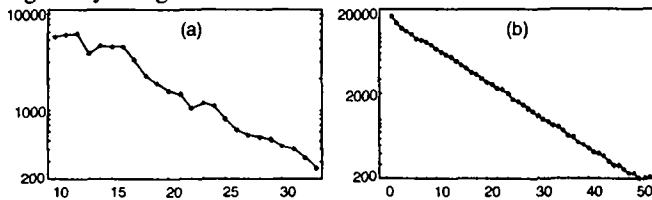
Figure 3.13 Cut-off frequency (Hz) versus tap number for a) the silicon cochlea by Watts, et al. and b) the improved version.

3.2.10 Discussion

In the above section we have seen that the silicon cochlea using CLBTs offers an improvement over previously realised silicon cochlea. Because of the reduction of mismatch in the bias currents of the different filter sections, the maximum gain is almost the same for each output and spacing of the cut-off frequencies is almost perfectly exponential. In Figure 3.14 we compare the

shape of the biological filter functions with the ones from the silicon cochlea. The basilar membrane displacement in the squirrel monkey was measured by Rhode (1991) at two points, 1.5 mm apart on the basilar membrane. The gain is the ratio between the basilar membrane displacement and malleus displacement. The two sets of points are the actual measurements, and the curves are the best fits as calculated by Rhode.

Figure 3.14 a) Filter gain in the squirrel monkey at 2 points (Rhode, 1971), and b) gain of the silicon cochlea at 5 different output taps.

Since the scale factor in both graphs is the same, the curves from the silicon cochlea should ideally have the same form. The main difference between the electronic and biological curves lies in the region of the peak gain, the silicon-cochlea curve being much wider 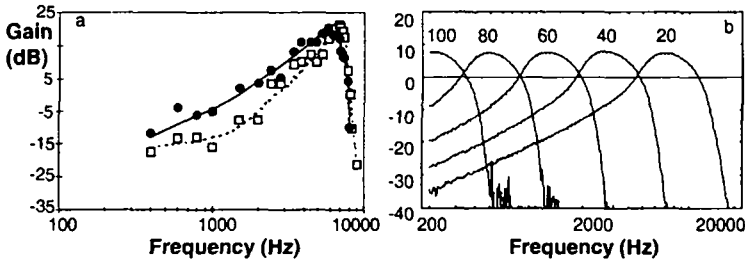around this point. A higher Q factor setting for the cochlear filters would make the filters more selective, but the gain would quickly increase to very large values, and the internal noise accumulated over all the previous sections in the filter cascade will be amplified to values that saturate the transconductance amplifiers in the filters so that the filters do not function properly anymore. It is also worth noting that the biological curves were measured at an intensity of 80 dB SPL, which is quite high; in Figure 2.10 we have seen that this yields response curves with a low selectivity. So one output of the silicon cochlea is even less selective than a point on the basilar membrane of a biological cochlea when the latter is at its least selective.

Another problem with the use of the current filter cascade, which is not shown in the above measurements, is that the delay between the output signal at a given filter stage and the input signal depends on the number of stages before the current stage. This means that if we increase the resolution, i.e., the number of filters in a given frequency range, the delay accumulated over this frequency range increases. In the biological cochlea, the delay over the auditory frequency range is fixed and is about 5 ms at the apex of the cochlea (von Békésy, 1960), which limits the maximum number of filters we can have in the cascade.

Notwithstanding these disadvantages of the current implementation, the second-order low-pass filter cascade offers a simple way to implement

exponentially scaled filters that approximate the shape of the auditory filters, with their shallow low-frequency slope and their steep high-frequency cut-off.

3.3 The inner hair cell model

3.3.1 Modelling the static inner hair cell voltage response

Although the fine details of the operation of the inner hair cell are still not fully understood, we know enough of the input/output relation of the inner hair cell to construct an electrical model. Figure 2.12 shows us that the relation between bundle deflection and the percentage of open ion channels has a sigmoidal form with a certain offset, so that 20% of the channels are open at equilibrium.

This can be easily modelled using a differential pair (see Figure 3.15) for which the current I_{sig} through transistor T_a is a sigmoidal function between 0 and I_{max} of the input voltage $V_{in} = V_{sig} - V_{ref}$, as we will see in the measurement of Figure 3.18. By using a differential pair with one transistor (T_b) four times as large as the other (T_a), only 20% of the bias current will flow through the smaller transistor (T_a) when V_{in} equals zero.
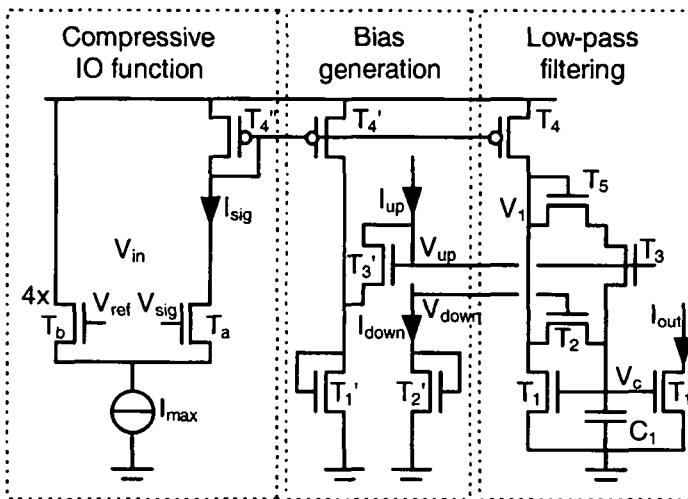

Figure 3.15 Circuit model of the frequency dependence of the intracellular voltage changes on the stimulus frequency.

3.3.2 Modelling the dynamic inner hair cell voltage response

The inner hair cell not only functions as a saturating non-linearity. As we can see in Figure 2.13, the inner hair cell also functions as a low-pass filter with a

cut-off frequency of about 1kHz. This means that we will have to create a low-pass filter with a time constant equal to $\tau = 1/2\pi f = 0.16$ ms to model this effect.

Obtaining such large time constants is one of the main problems in modelling comparatively slow brain elements with analogue VLSI circuits. If the linearity of the filter is not an important issue, then large 'time constants' can be realised using the current mirror shown in the low-pass filter blocks in Figure 3.15. This current mirror creates a non-linear low-pass filter for which I_{up} controls the maximum rise speed and I_{down} sets the maximum fall speed of the voltage on the capacitor C_1. We will take a closer look at the behaviour of this non-linear low-pass filter in the following section.

3.3.3 Analysis of the non-linear low-pass filter

It is complicated and not very informative to analyse such a non-linear filter completely, but we may obtain an intuitive understanding of the behaviour of the circuit by looking at its response to a large increase and a large decrease in the signal current I_{sig}. 'Large' in this context means a change in I_{sig} larger than either I_{up} or I_{down}. Since we want large time constants, I_{up} and I_{down} will be very small, so that any change in I_{sig} will be large by default.

With an increase in I_{sig} larger than I_{up}, the voltage V_1 will quickly rise to a voltage close to the positive power supply. This will make the diode-connected transistor T_5 in series with T_3 conduct, which raises the drain voltage of T_3, so that T_3 is saturated. We assume that the starting level of I_{sig} was higher than I_{down}, and transistors with a common gate connection in Figure 3.15 have the same device geometry. In this case the difference in source and gate voltages will be small for T_2, so that we can ignore its contribution. We can then write for the voltage V_c on the capacitor:

$$C\frac{dV_c}{dt} = I_{S3}\, e^{\frac{V_{up}\,-\,V_{T0}\,-\,n\,V_c}{n\,U_T}} = I_{S3}\, e^{\frac{V_{up}\,-\,V_{T0}}{n\,U_T}}\, e^{\frac{-V_c}{U_T}} \tag{21}$$

with I_{S3} defined as I_S in equation (2) applied to transistor T_3.

The bias voltage V_{up} is generated using I_{up} and a copy of I_{sig}, which ensures that V_{up} adapts to the signal level. With V_{up} independent of the signal current, there would be a limit to the capacitor voltage V_c above which C_1 can not be charged. Furthermore, the maximum current through T_3 would then depend strongly on the level of I_{sig}.

With I_{sig} much larger than I_{up}, we can simplify the calculation of the bias generation, assuming only I_{sig} going through T_1' and only I_{up} going through T_3'. Also, because V_{up} will change much faster with a change in I_{sig} than V_c, we will assume the change in V_{up} to be instantaneous. We can then write for V_{up}:

$$I_{S3} e^{\frac{V_{up} - V_{T0}}{n U_T}} = I_{up} \left(\frac{I_{sig}}{I_{S1}}\right)^n e^{\frac{V_{T0}}{U_T}} \tag{22}$$

and substituting (22) in (21) yields:

$$C \frac{dV_c}{dt} = I_{up} \left(\frac{I_{sig}}{I_{S1}}\right)^n e^{\frac{V_{T0} - V_c}{U_T}} \tag{23}$$

which we can rewrite as:

$$e^{\frac{V_c}{U_T}} dV_c = \frac{I_{up}}{C} \left(\frac{I_{sig}}{I_{S1}}\right)^n e^{\frac{V_{T0}}{U_T}} dt \tag{24}$$

and solve for V_c:

$$e^{\frac{V_c}{U_T}} = \frac{I_{up}}{C U_T} \left(\frac{I_{sig}}{I_{S1}}\right)^n e^{\frac{V_{T0}}{U_T}} t + e^{\frac{V_0}{U_T}} \tag{25}$$

where V_0 is the value of V_c immediately before the input signal change. With :

$$I_{out} = I_{S1} e^{\frac{V_c - V_{T0}}{n U_T}} \tag{26}$$

we obtain for the output current:

$$I_{out} = \left(\frac{t I_{up}}{C U_T}\right)^{1/n} I_{sig} + I_{out0} \tag{27}$$

where I_{out0} is the output current immediately before the input signal change.

The above equation is only valid as long as V_1 remains high enough to keep the transistors charging the capacitor saturated. When I_{out} approaches I_{sig}, V_1 will approach V_c and at equilibrium V_1 will be equal to V_c. As an approximation we can assume equation (27) to be valid until the moment I_{out} becomes equal to I_{sig}, after which it stays constant.

For a large decrease in I_{sig}, V_1 will decrease to a value close to the negative power supply. The diode-connected transistor prevents the capacitor from discharging through T_3. With V_1 so close to the negative power supply, the two transistors T_2' and T_2 form a approximate current mirror so that we can write for the voltage on the capacitor :

$$C \frac{dV_c}{dt} = I_{down} \tag{28}$$

which yields for I_{out} :

$$I_{out} = I_{out0} \, e^{\frac{-I_{down} \, t}{C \, n \, U_T}} \tag{29}$$

Again this equation is only a valid approximation as long as V_1 stays close to the negative power supply. However, to simplify this analysis, we will assume that (29) is valid until I_{out} becomes equal to I_{sig}, and that I_{out} is constant after that.
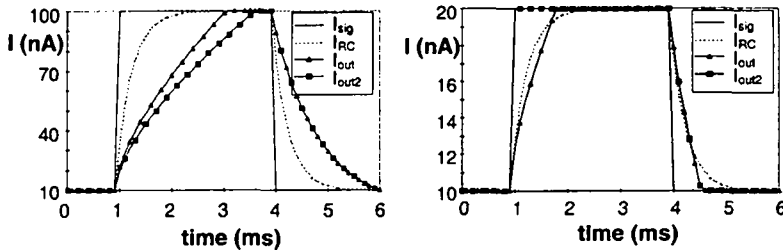

Figure 3.16 Step response of the low-pass filter block of Figure 3.15 compared with an ideal low-pass filter.

Figure 3.16 shows the results of equations (27) and (29) for two different input steps, 90 nA and 10 nA, when I_{sig} is 10 nA at rest. In these figures, I_{sig} is the input to the filter, I_{RC} the output of an ideal low-pass filter, and I_{out} the waveform obtained by applying equations (27) and (29). As we can see, this waveform is asymmetric and its 'time constant' isn't constant, but increases with signal level.

I_{out2} (Figure 3.16) is the response obtained by fixing V_{up} independently of the signal level to an arbitrary value chosen to obtain a slightly slower response to the 90 nA increase of I_{sig}. However this setting responds much too fast to a small positive change in I_{sig}, as shown by the response to a 10 nA step, where I_{out2} follows the rising flank of I_{sig} instantaneously. Therefore, although this solution is simpler, it can not be used when the amplitude of the input signal changes over an order of magnitude or more.

The response to the decreasing step of I_{sig} is also faster for the smaller step, but not by so much as to rule out the use of this solution.

3.3.4 Hair cell adaptation

Finally, Figure 2.14 shows that the conversion of the intracellular voltage into neuro-transmitter release shows adaptation to the stimulus. (In fact, Figure 2.14 shows a histogram of the spikes measured on a single auditory nerve fibre, but since the spiking probability of a spiral ganglion cell is directly proportional

to the probability of neuro-transmitter release by the inner hair cell it contacts, we can use this figure to model neuro-transmitter release.)
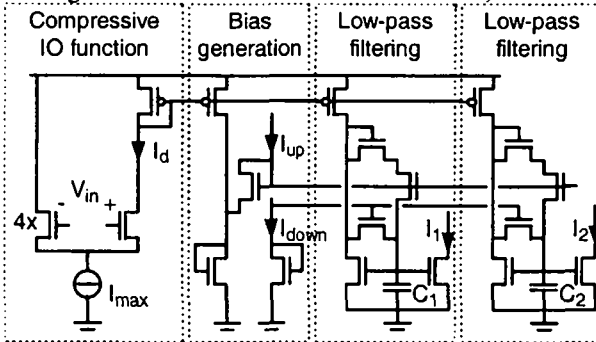

Figure 3.17 An electronic inner hair cell model

A form of adaptation similar to the one shown in Figure 2.14 can be obtained by taking the difference between I_1 and a second low-pass filtered version of I_{sig}. The same structure as shown in Figure 3.15 can be used to create the second low-pass filtered version of I_{sig}. When we use a larger capacitor in the second low-pass filter, I_2 will react more slowly to an onset of I_{sig} than I_1 does. In the circuit I have used $C_2 = 3C_1$. The final circuit modelling the input/output relation of the inner hair cell is given in Figure 3.17. The actual output current will be created with an additional current mirror, not shown in Figure 3.17 for clarity. This yields $I_{out} = I_1 - AI_2 + I_{spont}$, where A is the gain of the mirror and controls the ratio between the peak response and the sustained response of the circuit. A should be smaller than one in order to keep a sustained response to the input signal. I_{spont} sets the 'spontaneous rate of firing' of the inner hair cell model.

3.3.5 Test results

The four inner hair cell circuits have been realised in ES2's ECPD10 (1.0 μm) technology on a 1 mm^2 die. All parameters are shared between the four IHC circuits, but they have separate inputs and outputs.

Figure 3.18 (right) shows the measurements of the current I_1, which is the equivalent of the intracellular voltage in the real inner hair cell, as a function of the input voltage. Comparison with the intracellular voltage response to hair displacement of the real inner hair cell shows good agreement between the two. Although the differential pair has been designed to have an output current equal to 20% of the maximum current at zero input voltage, the measurement shows an output current equal to 30% of the maximum current in this case. This is probably due to the mismatch of the differential pair transistors.

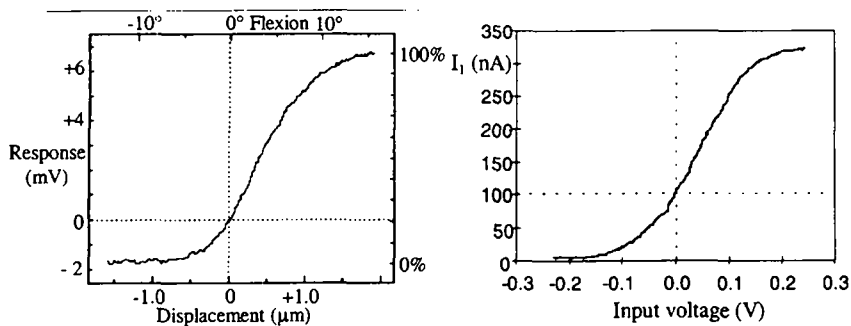**Figure 3.18** Measured voltage response to hair displacement of the inner hair cell (left) and measured output current as function of input voltage from the IHC circuit (right) for constant input signals.
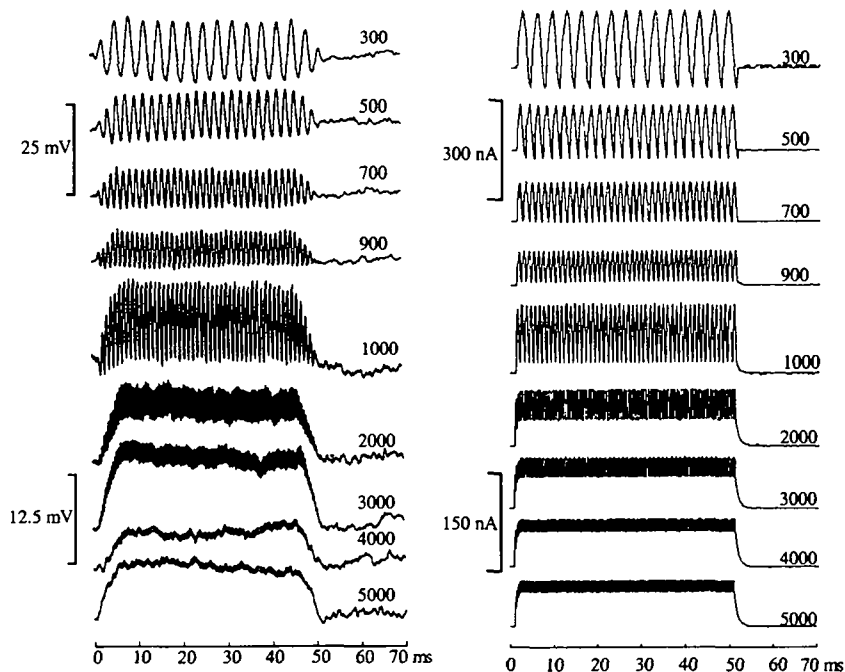

Figure 3.19 Intracellular voltage response of the inner hair cell (left) and measured I_1 changes from the IHC circuit for different frequencies of pure tone stimulation. Note the change of scale for the lower five traces.

Figure 3.19 compares the response of I_1 (measured before the adaptation stage) with the intracellular voltage response of an inner hair cell to pure tone stimulation at frequencies between 300 Hz and 5 kHz. The difference in the onset and offset of the response between the inner hair cell and the IHC circuit in Figure 3.19 is due to differences in the stimulus tone. In our measurements we used a tone burst with a square envelope, whereas Palmer and Russell (1986) used a stimulus with a gentle onset and offset.
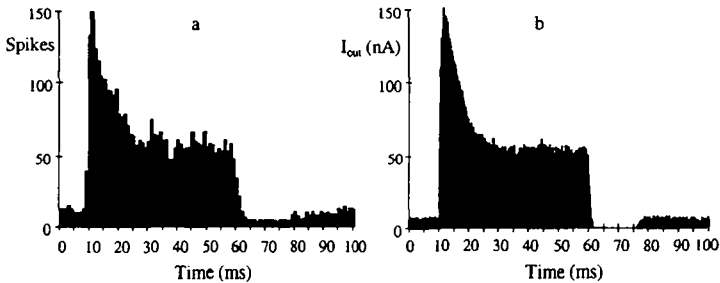
Figure 3.20 Auditory nerve fibre response (left) and the output of the electronic inner hair cell circuit after the adaptation stage (right).

The final output current of the circuit in is given by: $I_{out} = G(I_1 - AI_2 + I_{spont})$. Using I_{spont}, A, G, V_{up} and V_{down} allows us to create an output signal (Figure 3.20b) similar to the PSTH of a single auditory nerve (Figure 3.20a) when stimulated with a 5 kHz pure tone.

3.3.6 Discussion

Even though we can get good agreement between the measurements of the inner hair cell and of the IHC circuit, several discrepancies must be noted. These are related to the non-linearity of the low-pass filter, the fact that we use a continuous output current instead of spikes, and the manner in which adaptation is obtained.

Low-pass filtering

The fact that the time constant of the non-linear low-pass filter gets smaller for smaller changes in I_{sig} shows up in Figure 3.19 in the lower traces. In the 5 kHz trace, the a.c. component should be invisible, but because the time constant is short for a very small input change, the time constant will also be short for the first small part of a very large input change. This makes the a.c. component larger than it should be for the high frequency traces, but this has negligible effect on the spike timing of the spiking neurons for which this IHC signal is the input, as we will see in later chapters.

The fact that the match between the receptor potential and I_1 depends on the level of the input current and thus ultimately on sound intensity is another discrepancy, also related to the fact that the real inner hair cell behaves more like a linear low-pass filter than the IHC circuit.

A solution to these problems might be obtained by the use of so-called log-domain filters (Frey, 1993). These allow low-pass filters to be implemented in the current domain with only a few transistors; the filters will stay linear as long as the MOS transistors stay in weak inversion, or over a much larger range when bipolar transistors are used. I have discovered the possibilities offered by these circuits too late to implement and test the behaviour of such filters, and therefore I shall not present them here. However, if they work as well in practice as in theory, they should correct the sound intensity dependence of the current IHC circuit. Moreover, they would also allow the differential equations of existing models, such as Meddis' inner hair cell model, to be directly mapped onto silicon.

Spikes vs. current output

Figure 3.20 shows that the output current of the IHC circuit to a 5 kHz tone burst can be made to resemble the average number of spikes of an auditory nerve fibre after multiple stimulations at this frequency. However, the output current of the circuit is a continuous variable, whereas a spike is either present or absent. When a packet of neurotransmitter is released into the synaptic gap between the inner hair cell and a spiral ganglion cell, the ganglion cell almost always generates an auditory nerve spike. The number of AN spikes at any given time in Figure 3.20 is thus proportional to the instantaneous probability of transmitter release, which is a continuous variable. We can therefore use the output current of the IHC circuit as an analogue of the instantaneous spike probability. Since about ten spiral ganglion cells synapse with a single inner hair cell, and neighbouring inner hair cells receive almost the same stimulus from the basilar membrane, an alternative way to think of the circuit's output current is as an analogue of the average spiking rate of a group auditory nerve fibres that contact inner hair cells along a small zone along the basilar membrane. In chapter 5 we will return to the relation between the IHC circuit's output and the spikes on the auditory nerve, when we will compare the spiking behaviour of the spiking neuron circuit with some neurons in the cochlear nucleus which are innervated by the auditory nerve.

Adaptation

Although subtracting the two low-pass filtered currents as in Figure 3.20b yields an adaptation that looks similar to the adaptation due to the depletion of available neuron transmitter as in Figure 3.20a, there is a basic difference

between the two adaptations. The reduction of available neurotransmitter changes the gain between receptor potential and the amount of neurotransmitter released and is thus a multiplicative process, and not a subtractive process as in the circuit. This is visible in Figure 3.20 after the off-set of the tone burst. Whereas the gain of the inner hair cell remains reduced and slowly recovers, the value of $I_{out} = G(I_1 - AI_2 + I_{spont})$ becomes negative when I_{spont} is small. Because the actual output current of the circuit passes through a current mirror, which truncates the signal to values larger than or equal to zero, we do not see this clearly in Figure 3.20b. Consequently, there is only a small difference between the reaction of a real inner hair cell and the circuit to the offset of tone burst.

The situation is different, however, when the inner hair cell is stimulated with a continuous high frequency tone which is stepped up and down in amplitude. We can see in Figure 3.20a that after adaptation the gain between the input of the inner hair cell and the spike rate on the auditory nerve is only a third of what it was at the onset. This is shown by the fact that the onset spike rate is three times higher than the sustained rate. Now let us consider the case where 50 ms after the onset of the tone the amplitude of the tone is increased again by a factor of three instead of being switched off. At first this will increase the spike rate by a factor of three, i.e., again to 150 spikes per bin width, but a new adaptation will take place so that the sustained rate will be lower. However, the sustained rate for the higher intensity sound will remain higher than the sustained rate for the lower intensity. Let us assume that the gain adapts from one-third to one-fifth. The sustained rate will then become 90 spikes per bin width for the second tone. When we now reduce the amplitude of the tone by a factor of three to its original value, the spike rate will first reduce by a factor of three to 30 spikes per bin width, and then slowly adapt back to a sustained rate of 50 spikes per bin width. I have schematically drawn this in Figure 3.21a.
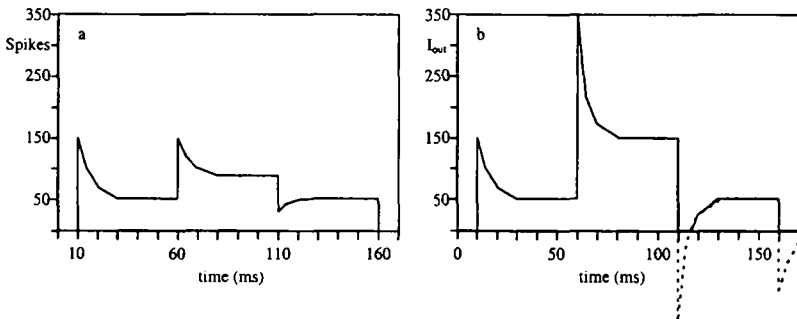

Figure 3.21 Schematic representation of the response of an inner hair cell (a) and the IHC circuit (b) to changes in amplitude of the stimulus tone.

The inner hair cell circuit will behave differently. The output current of the circuit is given by: $I_{out} = G(I_1 - AI_2 + I_{spont})$ and I_{spont} is almost zero. If we assume for the sake of simplicity that I_1 follows the change in the input signal instantaneously, and only I_2 is a low-pass filtered signal, it is easy to see that the adaptation gain A has to be 2/3 in order to obtain a sustained output of 50 nA when the peak output is 150 nA as in Figure 3.20b. In reality I_1 is also low-pass filtered but with a time constant shorter than I_2; I_1 doesn't change instantaneously, but still a lot faster than I_2. Ignoring this, we can write in the adapted situation: $GI_1 = 150$ nA and $GAI_2 = 100$ nA. Both G and A are constants which can be modified externally, but do not vary as a function of the input signal. Increasing the input amplitude by a factor of three, 50 ms after the onset of the tone, will thus yield a peak current of $3*150-100 = 350$ nA, and I_2 will then start to increase to a three times higher value, so that the sustained rate becomes 150 nA. Reducing the input amplitude by a factor of three will make GI_1 equal to 150 nA again, and GAI_2 will stay at first at 300 nA so that I_{out} should become -150 nA, but since it is limited to positive values, I_{out} reduces to 0 nA.

I've drawn this behaviour in Figure 3.21b. Of course the assumption that I_1 responds instantaneously exaggerates the onset and offset response of the IHC circuit, but it serves its purpose in showing clearly how the IHC circuit's response — a subtractive process — differs from the response of an inner hair cell — a multiplicative process.

This problem with the IHC circuit will have to be corrected if we want to build an electronic hearing system which behaves like the human auditory system in every situation. However, in the more limited experiments currently performed in the domains of auditory physiology and psycho-acoustics, relatively simple stimuli are used, and the IHC circuit can be set to respond nearly correctly for a given stimulus. One of the stimuli often used is a short burst of sound, like in Figure 3.20, for which the adaptation can be set correctly. Other stimuli used are those with relatively long durations, for which the onset and offset responses are not important because most information is taken from the sustained rate. This is for instance the case in the pitch detection experiments presented in chapter 7, for which amplitude modulated continuous pure tones are used. Although the modulation of the amplitude at very low frequencies has an effect on the adaptation by the inner hair cell, this effect is negligible for modulation frequencies in the 100 Hz range. In these experiments we can thus switch the adaptation in the IHC circuit off ($A = 0$), and adapt the gain G to reflect the gain in the sustained situation.

3.4 Summary

In this chapter, I have presented a working analogue VLSI model of the basilar membrane. This silicon cochlea has been tested and showed a good regularity in its frequency gain curves and the spacing of its cut-off frequency. Even though the silicon cochlea is much less-frequency selective than the biological cochlea, even at high sound intensities, it still offers a relatively simple way to model the exponential filtering of the basilar membrane.

I have also presented an electronic inner hair cell model. This IHC circuit has also been tested and measurements of the equivalent of the intracellular voltage of the inner hair cell shows good agreement with measured voltages in the biological inner hair cell. Furthermore, an adaptation similar to the adaptation seen on the auditory nerve can be obtained.

3.5 References for chapter 3

Arreguit, X. (1989) *Compatible Lateral Bipolar Transistors in CMOS Technology : Model and Applications*, Ph.D. thesis, Ecole Polytechnique Fédérale, Lausanne.

Fragnière, E., van Schaik, A., and Vittoz, E.A. (1997) "An analogue VLSI model of active cochlea." in *Analog Integrated Circuits and Signal Processing*, Kluwer Academic Publishers, Vol. 13, No. 1/2, pp. 19-36.

Frey, D.R. (1993) "Log-domain filtering: An approach to current-mode filtering," *IEE Proceedings G*, vol. 140, pp. 406-416.

Lyon, R.F. and Mead, C.A. (1988) "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1119-1134.

Lyon, R.F. (1991) "Analog implementations of auditory models," *Proc. DARPA Workshop Speech and Natural Language.* San Mateo, CA:Morgan Kaufmann.

Liu, W. et. al. (1991) "Analog VLSI implementation of an auditory periphery model," *Advances Res. VLSI, Proc. 1991 Santa Cruz Conf.*, MIT Press, pp. 153-163.

Mead, C.A. (1989) *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, MA.

Palmer, A.R. and Russell, I.J. (1986) "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells." *Hearing Research*, Vol. 24, pp. 1-15.

Sarpeshkar, R., Lyon, R.F., and Mead, C.A., (1996) "An analog VLSI cochlea with new transconductance amplifiers and nonlinear gain control," in *Proc. IEEE International Conference on Circuits and Systems*, Atlanta, Vol 3, pp. 292-295.

Vittoz, E. (1983) "MOS transistors operated in the lateral bipolar mode and their application in CMOS technology," *IEEE J. Solid-State Circuits*, vol. SC-24, pp. 273-279.

Vittoz, E. (1985) "The design of high-performance analog circuits on digital CMOS chips," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 657-665.

Watts, L., Kerns, D.A., Lyon, R.F. and Mead, C.A. (1992) "Improved implementation of the silicon cochlea," *IEEE J. Solid-State Circuits*, vol. SC-27, pp. 692-700.

Watts, L. (1993) *Cochlear Mechanics: Analysis and Analog VLSI.* Ph.D. thesis, California Institute of Technology, Pasadena.

4. The Spiking Neuron

4.1 Introduction

In the previous chapter I have shown that it is possible to simulate the auditory periphery using analogue VLSI hardware with a certain amount of fidelity. While further design improvements are certainly possible, we can say that it is now possible to simulate the response of the human ear using electronic circuitry. This will provide the necessary basis for extending the scope of VLSI circuits into the auditory pathway, in which the bulk of the signal processing is performed by spiking neurons (Figure 4.1).

Auditory Cortex
Medial Geniculate Body
Inferior Colliculus and
Lateral Lemniscus Nuclei
Cochlear Nucleus
Superior Olivary Complex
Cochlea
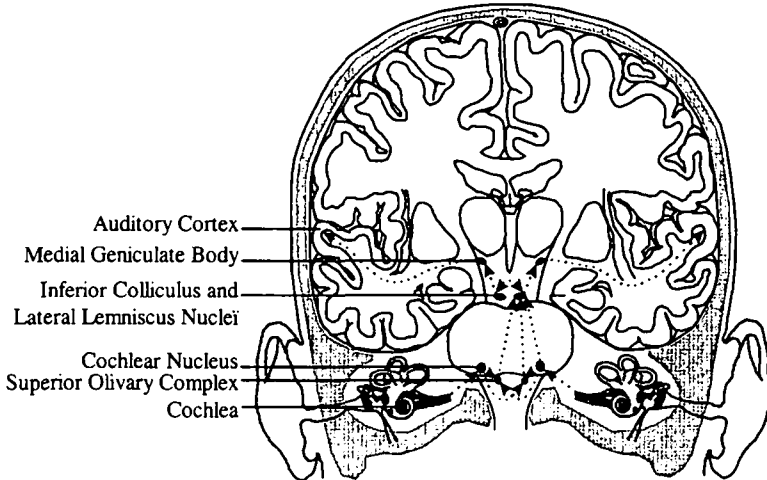
Figure 4.1 Simplified diagram of the auditory pathway. The projections to the Inferior Colliculus on the left are omitted for clarity.

Before we can start creating electronic models of neural circuits in the auditory pathway, we will therefore have to understand and model these spiking neurons. In this chapter we will look at the electrophysiology and anatomy underlying the spiking behaviour, and in the next chapter I will develop an

electronic model of this spiking behaviour. Most information on the operation
of the spiking neuron in this chapter is taken from Kandel, et al. (1991)
and Alberts, et al. (1983).

4.2 Passive membrane properties

Figure 4.2 shows a drawing of a typical nerve cell, or neuron. Although they
come in many different shapes and sizes, most neurons possess the different
features shown in this figure. The neuron receives input from other nerve cells
by synapses on its dendrites and on its soma (cell body), and the output of a
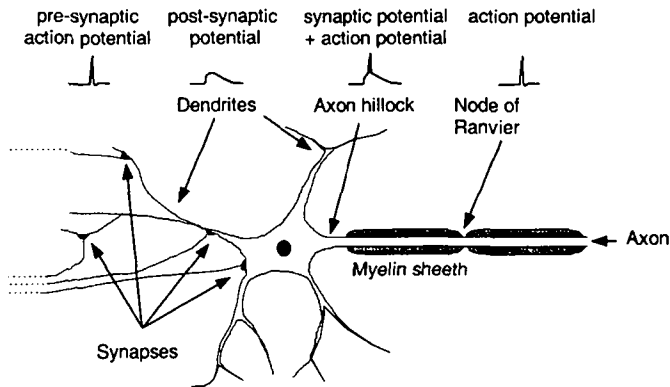neuron travels as a spike on its axon to other neurons.

Figure 4.2 A typical nerve cell and the potentials at the different sites

The inside (cytoplasm) of the neuron is separated from the extracellular fluid
by the cell membrane, which envelops the whole neuron. Like all other cells in
the body, neurons expend about 30% of their metabolic energy in pumping
sodium ions out of the cell and potassium ions into the cell. The cell membrane
contains ion pumps, embedded enzymes that transfer 3 Na^+ ions out of the cell
for every 2 K^+ ions pumped into the cell. This charges the interior of the cell
negatively and creates a concentration gradient for these ions, with a high
potassium concentration inside the cell and a high sodium concentration
outside. These gradients serve to actively transport other molecules into or out
of the cell by chemically binding to these ions. Most cells in the body have
membranes that are mainly permeable to potassium; the concentration gradient
causes potassium ions to leak out through ion channels selective to K^+. This
charges the cell even more negatively with respect to the extracellular fluid,
which in turn reduces the flow of the positively charged potassium ions. At the
same time, sodium slowly leaks into the cell through sodium channels in the

membrane. Equilibrium is reached when the passive currents of both ions through the cell membrane are counter-balanced by the Na^+-K^+ pump.

A third ion influencing the resting potential of the neuron is chloride. All neurons contain chloride channels. Because the chloride concentration is largest outside the cell, Cl^- leaks into the cell until the electrical force generated by the membrane potential counter-balances the chemical force created by the concentration gradient and no more Cl^- ions pass the membrane. This equilibrium potential for an ion is called the Nernst potential, after the German chemist Walter Nernst, and can be calculated with the following formula:

$$E_{ion} = Z\, U_T \ln\left(\frac{[ion]_o}{[ion]_i}\right) \tag{30}$$

where Z is the valence of the ion, U_T is the thermal voltage kT/q (27 mV at body temperature), and $[ion]_o$ and $[ion]_i$ are the ion concentrations outside and inside the cell. In neurons that do not have a Cl^- pump in addition to the Na^+-K^+ pump, the Nernst potential for the chloride ions E_{Cl} will be equal to the resting potential of the cell, E_R. Some neurons do have a chloride pump that actively pumps out chloride, which increases $[Cl^-]_o/[Cl^-]_i$ and thus makes E_{Cl} more negative than E_R.
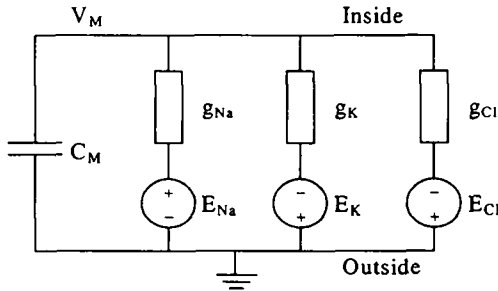
Figure 4.3 Equivalent electrical circuit for the membrane at rest.

It is the local charge separation across the cell membrane that creates the local membrane potential. Changes in membrane potential during activity of the neuron are thus created by changes in local charge separation across the membrane, not by changes in bulk concentration. For normal operation we can therefore consider the concentrations of the different ions inside and outside the cell to be constant and use them to calculate the Nernst potentials for each ion. In the nerve cells in the brain the Nernst potentials of the three ions are approximately: E_K=-75 mV, E_{Na}=+55 mV, and E_{Cl}=-60 mV. Whenever the membrane potential V_M is moved away from the Nernst potential for a given ion, the electromotive force will drive the ions into or out of the cell, depending

on the direction of change of V_M. The amount of ions crossing the cell membrane is directly proportional to the potential change and to the permeability of the membrane for that ion, which is proportional to the number of ion channels for that ion in the membrane. This ohmic relation can be modelled with an equivalent electrical circuit in which conductances are connected in series with batteries representing the Nernst potentials of the ions, as shown in Figure 4.3.

4.3 Action potential generation

Apart from the ion channels which are always open, the nerve cell also has ion channels whose conductance of a particular ion depends on the membrane voltage. These voltage-dependent conductances control the generation of the action potential (spike). The simplest neurons have only one voltage-dependent type of ion channel for potassium and one for sodium, but more complex neurons, such as the fusiform cell in the dorsal cochlear nucleus of the central nervous system, have at least three types of voltage-dependent potassium channels and two types of voltage-dependent sodium channels. We can incorporate these voltage-dependent conductances in the equivalent electrical circuit. Since the passive conductances of Figure 4.3 do not change, we can simplify the circuit by lumping these together and connecting them to a battery with a value equal to the resting potential of the neuron. In the simplest case this yields the circuit of Figure 4.4.

A model as in Figure 4.4 was originally developed by Hodgkin and Huxley (1952) to describe the generation of action potentials in the squid giant axon. According to their model, the following sequence of events takes place: A depolarisation of the membrane, i.e., an increase in V_M, causes rapid opening of sodium channels, and sodium is thus driven into the cell following its concentration gradient. The influx of sodium depolarises the membrane further, causing more sodium channels to open and resulting in even more sodium entering the cell. This positive feedback generates the action potential, which causes the membrane potential to rise towards E_{Na}. Two factors limit the duration of the action potential. (1) A second, slower process gradually inactivates the sodium channels, which stops the sodium influx. (2) Depolarisation of the cell also opens voltage-gated potassium channels with some delay, which creates an outflow of potassium ions. The first factor alone would be sufficient to slowly return the membrane potential to the resting voltage, due to the leakage of sodium ions across the cell membrane. However, the opening of the potassium channels increases the speed of repolarisation of the cell by a large factor. The sodium channels stay inactivated and the potassium channels remain activated for a while after an action potential. This causes the so-called refractory period of the neuron, divided in two parts: the

absolute refractory period during which it is impossible to generate a new action potential because the sodium channels are still inactivated, and the relative refractory period, during which a stronger than normal stimulus is needed to generate an action potential, because the potassium channels are still open after the sodium channels de-inactivate.
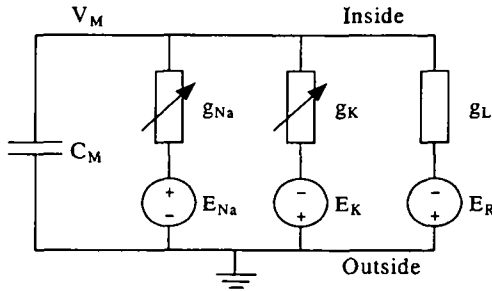

Figure 4.4 Equivalent electrical circuit including two voltage-dependent conductances.

Although the voltage-dependent conductances always change in a strictly graded manner, this model still exhibits a threshold-type behaviour for the generation of action potentials. A depolarisation will create a net influx of sodium, but will also increase the outward potassium current, and an outward leakage current through the passive ion channels. If the depolarisation is below threshold, the inward sodium current will be counterbalanced by the outward currents. However, the great voltage sensitivity of the sodium channels, together with the rapid kinetics of these channels, ensure that the inward sodium current exceeds the outward currents from the threshold voltage onward, and an action potential will be generated.

Other voltage-dependent conductances can be readily included in this model. They will alter the reaction of the cell to a given stimulus, thereby creating some of the more complex spiking behaviours.

4.4 Leaky-integrate-and-fire neuron model

Although the Hodgkin-Huxley model presented in the previous chapter is excellent for gaining insight in the way neurons generate spikes, it is not suited for implementation on silicon, since it will yield very large neuron circuits due the complexity of the Hodgkin-Huxley model. In their model, each voltage controlled conductance is governed by a set of equations. For instance, the equations for the sodium channels are:

$$g_{Na} = g_{NaMax}m^3h,$$

$$\frac{dm}{dt} = \alpha_m(1-m) - \beta_m m, \qquad \frac{dh}{dt} = \alpha_h(1-h) - \beta_m h$$

$$\alpha_m = \frac{25 - V_M}{10(e^{25-V_M/10} - 1)}, \qquad \beta_m = 4e^{-V_M/18}$$

$$\alpha_h = \frac{7}{100}e^{-V_M/20}, \qquad \beta_h = \frac{1}{e^{30-V_M/10} + 1} \qquad (31)$$

where V_M is the membrane potential normalised to 1 mV and g_{NaMax} is the maximum sodium conductance, m is the sodium activation variable and h is the sodium inactivation variable, both dimensionless quantities between 0 and 1. The values in the expressions for α_m, β_m, α_h, and β_h where determined empirically using curve-fitting for the squid giant axon by Hodgkin and Huxley. Other voltage-dependent membrane conductances can be modelled with similar sets of equations.

A much simpler way to capture the essentials of neural spike generation is used in the *leaky-integrate-and-fire (LIF) neuron model (Lapique, 1907)*. In this model only the passive membrane properties of the Hodgkin-Huxley model are kept, i.e., the 'leaky integrator' consisting of the membrane capacitance and the leakage conductance. Instead of modelling the voltage-dependent conductances, the model introduces an explicit spiking threshold for the membrane potential. We can describe this model with the following equations:

$$C_M\frac{dV_M(t)}{dt} = i(t) - g_L V_M(t), \quad V_M(t) < V_\theta \qquad (32)$$

If $V_M(t) = V_\theta$: set $t_s = t$ and generate a spike,

$$\text{at } t = t_s + T_S : \text{set } V_M(t) = 0, \qquad (33)$$

where i(t) is the input current, C_M the membrane capacitance, g_L the leakage conductance, $V_M(t)$ the membrane voltage, V_θ the threshold voltage, T_S the width of the spike, and t_s is the time t when $V_M(t)$ reaches V_θ. Equation (32) describes the leaky-integration of the input current on the membrane capacitance. In the absence of an input current, V_M will return to zero due to the leakage conductance. Equation (33) describes the spike generation process.

Whenever V_M reaches the threshold V_θ, a spike is generated and after a short delay T_S V_M is reset to its resting value. Normally T_S is considered zero in the LIF model, but since in the next chapter we will be dealing with a real circuit for which it is impossible to generate a spike of zero width, I will explicitly include T_S in the model here.

With a constant input current i, the LIF neuron will need a constant time T_I after the reset operation to reach the spiking threshold. Integrating (32) and solving for $V_M = V_\theta$ yields:

$$T_I = \frac{C_M}{g_L} \ln\left(\frac{i}{i - g_L V_\theta}\right) \tag{34}$$

The LIF neuron will thus spike regularly, for a constant input current, with a frequency given by:

$$f_{spike} = \frac{1}{T_I + T_S} \tag{35}$$

The above model does not model the refractory period of a neuron, although this is an important element in the spike generation process. It is however straightforward to add this to the model by introducing a time T_R after the generation of a spike during which the membrane potential is forced to zero (see for instance Tuckwell, 1988). In this case the spike frequency for a constant input current is given by:

$$f_{spike} = \frac{1}{T_I + T_S + T_R} \tag{36}$$

We can furthermore force the membrane potential of the LIF neuron to V_S whenever a spike is generated. This serves no functional purpose, but is a simple way to model the evolution of V_M at the axon hillock of a neuron. This evolution is shown in Figure 4.5 when the neuron is stimulated with a constant input current.
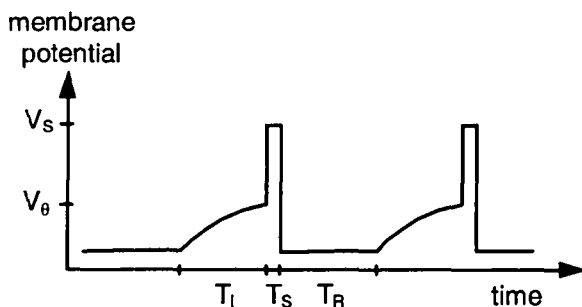
Figure 4.5 Typical evolution of the membrane potential in the modified LIF neuron for a constant input current.

4.5 Interactions between neurons

Neurons send their spikes to other neurons over their axon and contact other neurons through synapses. Longer axons are myelinated, which increases the distance between the intracellular fluid in the axon and the extracellular fluid, thereby reducing the capacitance of the axon. This increases the transmission speed of the action potential along the axon. The myelin sheet is furthermore interrupted at regular distances along the axon, at the so-called "nodes of Ranvier", to allow regeneration of the action potential by the opening of sodium channels in the membrane.

Synapses come in many different shapes and sizes, and the details of their operation vary widely. Some synapses are entirely electrical, but most synapses release a chemical substance (neurotransmitter) into the cleft between the pre-synaptic and the post-synaptic membrane upon arrival of a spike at the pre-synaptic terminal. This neurotransmitter binds with special receptors in the post-synaptic cell, causing specific ion channels to open. The type of neurotransmitter used depends on the type of synapse and its effect depends on the type of receptor.

The most common neurotransmitter in the brain is glutamate, which is always excitatory; at least four different classes of glutamate receptors have been identified however. The end effect of glutamate release is the opening of ion channels which are permeable both to sodium and potassium ions, and sometimes to calcium ions. Because the channels are more permeable to sodium than to potassium, their effect on the post-synaptic cell is still a net influx of positively charged ions. This creates an Excitatory Post-Synaptic Potential (EPSP).

The principal inhibitory neuro-transmitters in the brain are GABA and glycine, which open Cl⁻ channels in the post-synaptic cell. If the Nernst potential for chloride is lower than the resting potential of the post-synaptic

neuron, then the opening of chloride channels will hyperpolarise the cell and create an Inhibitory Post-Synaptic Potential (IPSP). However, if the post-synaptic cell doesn't contain a Cl⁻ pump, the Nernst potential for chloride will be the same as the resting potential of the cell and no IPSP will be created. The increased chloride conductance will make it harder for excitatory synapses in the neighbourhood to increase the membrane potential, resulting in so-called shunting inhibition.

It takes time for neurotransmitter released from a pre-synaptic cell to bind to the different ion channels. This introduces a delay, typically 1 ms or more, between the arrival of the pre-synaptic spike and the opening of the ion channels. Once the transmitter is bound to the channel it will have to be removed chemically, which will also take a certain amount of time, so that the ion channels will stay open for a short duration. Once the ion channels are closed again, the neuron will return to its resting state; the time course will depend on the time constant of the neuron.

Both the dendrites and the soma are relatively poor in voltage-dependent sodium channels. Potential changes at these sites will not create an action potential, but are passively transmitted along the cell membrane. At the start of the axon — the axon hillock — the voltage-dependent sodium channels become numerous, and it is here that the action potential is generated and transmitted along the axon.

Up to now, we have considered the neuron to be without internal resistance, so that any charge added to the membrane instantaneously increases the membrane potential at the axon hillock, where the spike is generated. Although this assumption is reasonable for post-synaptic potentials from synapses on the soma, the small diameter of the dendrites will introduce a resistance in the pathway to the cell body. Thus compared to an excitation on the soma itself, an excitation at a distal dendrite will create a smaller but longer lasting variation of the membrane potential measured at the soma. In the same way, an inhibitory synapse at a proximal dendrite will have a stronger influence on the excitation for more distal synapses on the same dendrite, than for the excitation at synapses on the soma. The neuron therefore does not merely add up excitatory and inhibitory inputs until the spiking threshold is reached at the axon hillock; it performs a spatial and temporal summation which depends critically on the passive properties of the cell, specifically its time and length constants.

4.6 Anatomical influences

From a signal processing point of view, spikes are the end result of a computation performed by the neuron on its input; they transmit this result to other neurons, possibly over long distances. A great deal of the computation performed by the neuron is determined by several anatomical factors.

Consider first the case of a larger cell. The size of the cell determines its membrane capacitance, which increases linearly with the membrane's surface. If the density of the passive ion channels in the membrane remains constant, then the leakage conductance g_L will also increase linearly with membrane area. The time constant of the cell, C_M/g_L, will then be independent of the cell's size. However, the total charge needed to reach the threshold voltage of the cell will increase with the cell size, resulting in a decreased relative strength of the individual synapses.

Next, suppose the number of leakage channels is reduced, increasing the time constant of the cell. Although the total charge needed on the membrane capacitor to reach the threshold voltage remains the same, each charge generated by an excitatory input will remain longer in the cell; this increases the time period in which several excitatory inputs can arrive at the cell and still create an action potential.

Finally, a large part of the computational function of a neuron is determined by its connectivity. If we know the different types of neurons from which a certain neuron gets its inputs, and if we know whether the synapses from these input neurons are excitatory or inhibitory, then we can make an educated guess about the operation of the neuron, based on the membrane properties of the cell. Physiologically, although the synapses and the membrane properties can be determined by studying a single cell, it is hard to trace back the synapses to determine where the inputs are coming from. In the auditory pathway, it is the cochlear nucleus that has been studied in most detail (see for instance Merchan, et. al., 1993). Its main source of input is known since most inputs to the neurons in the cochlear nucleus are from the auditory nerve (see Figure 4.1), which makes it a good site to test hypotheses about the influence of synapses and membrane properties on the computational performance of a neuron. In the next chapter we will present a silicon spiking neuron and use the knowledge gathered from physiological studies of the cochlear nucleus (CN) to model its neurons. Then we will compare the output of the circuit with the response of typical CN neurons in response to tone bursts.

4.7 Summary

In spiking neurons, the intracellular fluid is separated from the extracellular fluid by the cell membrane which contains selectively permeable ion channels. At rest, the membrane potential is negative and potassium is most concentrated inside the cell, whereas sodium is most concentrated outside the cell.

A spike is generated under the control of voltage-dependent ion channels at the axon hillock when the cell membrane is depolarised above a critical value. First, sodium channels open, which creates an influx of positive charge; this depolarises the cell even more, which opens more sodium channels. This

positive feedback creates the upswing of the spike. A second slower process inactivates the sodium channels, and allows the charge to be removed from the cell. The downswing of the spike is accelerated by the delayed opening of voltage-dependent potassium channels.

After a spike, the potassium channels stay activated for a while and the sodium channels inactivated. This creates a refractory period, during which it is impossible — or much more difficult — to stimulate the cell to create another spike.

Interactions between neurons take place at the synapses; the effect of a pre-synaptic spike can be either excitatory or inhibitory.

The potential at the axon hillock result from spatio-temporal summation of all excitatory and inhibitory inputs; the characteristics of this summation process depend critically on the passive properties of the nerve cell.

4.8 References for chapter 4

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. (1983) *Molecular Biology of the Cell*, Garland Publishing, Inc., New York & London.

Kandel, E.R., Schwartz, J.H., and Jessel, T.M. (1991) *Principles of Neural Science*, Prentice-Hall International, London.

Lapicque, L. (1907) "Recherches quantitatifs sur l'exitation électrique des nerfs traitée comme une polarisation," *J. Physiol.(Paris)*, Vol. 9, pp. 620-635.

Merchan, M.A., Juiz, J.M., Godfrey, D.A., and Mugaini, E. (1993) *The Mammalian Cochlear Nuclei: Organisation and Function*, Plenum Press, New York.

Oertel, D. (1983) "Synaptic responses and electrical properties of cells in brain slices of the mouse anteroventral cochlear nucleus," *Journal of Neuroscience*, Vol. 3, pp. 2043-2053.

Tuckwell, H.C. (1988) *Introduction to theoretical neurobiology*, Cambridge University Press, Cambridge UK.

5. The Silicon Neuron

5.1 Introduction

As we have seen in the previous chapter, much is known about the details of action potential generation. Various voltage-dependent ion channels influence the neural spiking behaviour. Although theoretically it might be possible to incorporate all known details into our electronic model, we shall have to make a trade-off between the detail incorporated in the model and the actual size of the circuit. One approach, started by Mahowald and Douglas (1991) is to implement neuron circuits which are analogue approximations of the Hodgkin and Huxley model. This can yield highly detailed single neuron models, but takes up a lot of area. For instance, a recent publication based on this approach (Rasche, et al., 1997) describes a silicon neuron that emulates a cortical pyramidal cell. They were able to put a single neuron featuring about 30 adjustable parameters on a 4 mm^2 chip. However, in most cases, building a single neuron model is not the ultimate goal, but only a necessary step in order to simulate the collective behaviour of a large group of neurons. It is thus important to minimise the amount of detail in order to reduce the circuit size, allowing more neurons to be put on the same chip.

5.2 An electronic spiking neuron model

5.2.1 Implementation

A very simple neuron model is shown in Figure 5.1. The membrane of a biological neuron is modelled by a membrane capacitance, C_{mem}; the membrane leakage current is controlled by the current source, I_{leak}. In the absence of any input (i=0), the membrane voltage will be drawn to its resting potential (controlled by V_-) by this leakage current. Excitatory inputs (i>0) simply add charge to the membrane capacitance. Inhibitory inputs are modelled by a negative input current (i<0). If an excitatory current larger than the leakage current is injected, the membrane potential will increase from its resting potential. This membrane potential, V_{mem}, is compared with a controllable threshold voltage V_{thres}, using a basic transconductance amplifier driving a high

impedance load. If V_{mem} exceeds V_{thres}, an action potential will be generated. I will discuss the generation of the action potential in detail in the next section.

In the biological neuron spikes are sent to other neurons over the axon of the neuron. In order to send the spike over reasonably long distances, the spike is regenerated on the axon at specialised nodes, called nodes of Ranvier. Although the membrane potential at the axon hillock changes slowly when it integrates an excitatory input current, spikes which are regenerated on the axon do not reflect this, and only the event of an action potential is transmitted. On-chip, we do not have to regenerate the spikes because the metal wires have little resistance, at least for the distances we are dealing with. We could thus in principle transmit the membrane potential V_{mem}, but in order to be able to use spike-based computation as in biological neural systems, we shall have to send spikes. However, we do not need to add a node of Ranvier model to the circuit in order to generate a fixed width, fixed height spike; we will see in section 5.2.2, we can simply use the spike shown in the top right of Figure 5.1 as the output of our model. This signal is inverted with respect to biological spikes, i.e., the spike goes downward. We could of course invert the spike and we will see in section 5.4.3 that we will need to do so to create inhibitory synapses, but the downward spike is directly compatible with excitatory synapses.

The above circuit is very similar to the sodium-potassium neuron circuit described by Sarpeshkar et al. (1992) in an internal report of the California Institute of Technology. Their implementation uses the current in the left branch of the comparator in Figure 5.1 to create the upswing of the spike and to charge the capacitor C_K. This allows them to use a few transistors less than in my implementation, but also results in less freedom to control the different parameters of the neuron independently. This circuit is only slightly smaller, since most area is taken up by the two capacitors anyway. They claim that reducing the number of parameters to adjust is also important to reduce the physical size of a network of neurons, since each parameter is routed by a wire of non-zero width. However, this is only important when each neuron has to receive its own set of parameters that have to be routed to each neuron individually. In most cases, a large number of their parameters will be shared, while other parameters vary in orderly fashion along the array. This means that each extra parameter only introduces one extra metal or resistive wire, or only a few transistors to control the parameters as in the dendritic tree circuit which we will discuss later in this chapter (shown in Figure 5.14). Again, the savings in area are small due to the relatively large capacitors, and I have preferred to keep the circuit more a flexible building block by introducing a few extra parameters. Sarpeshkar's circuit is a good and straightforward way of simplifying my circuit though when the extra options are not needed.

I have realised 32 copies of the neuron shown in Figure 5.1, together with some circuitry to facilitate communication of signals on- and off-chip (see section 5.4.1) on a 1 mm x 2.5 mm die, using the ECPD10 (1 µm) technology of ES2. All transistors have a W/L ratio of 10 µm/10 µm except for the switches and inverters, which are 2 µm/10 µm, and C_K and C_{mem} are 10 pF.
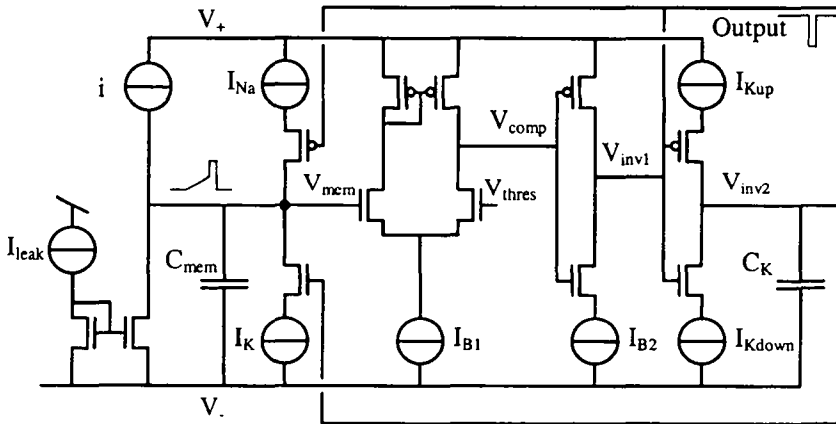

Figure 5.1 A simple electronic neuron circuit.

5.2.2 Spike generation

The generation of the action potential in the neuron circuit is patterned after the biological neuron, in which an increased sodium conductance creates the upswing of the spike and in which the delayed blocking of the sodium channels plus delayed increase of the potassium conductance creates the downswing. The circuit models this as follows: If V_{mem} rises above V_{thres}, the output voltage of the comparator V_{comp} will rise to the positive power supply. The output of the following inverter V_{inv1} will thus go low, thereby allowing the "sodium current" I_{Na} to pull up the membrane potential. At the same time, however, a second inverter will allow the capacitance C_K to be charged at a rate controlled by the current I_{Kup}. As soon as the voltage V_{inv2} on C_K is high enough to allow conduction of the NMOS whose gate it controls, the "potassium current" I_K will be able to discharge the membrane capacitance. Although the operation of the circuit differs slightly from the biological neuron model in that the potassium current only creates the downswing of the spike and that the sodium current only switches off again when V_{mem} drops below V_{thres}, the spiking behaviour of the neuron does not seem to be affected by this.

Two different potassium channel time constants govern the opening and closing of the potassium channels. The current I_{Kup} which charges C_K controls the spike width, since the delay between the opening of the sodium channels and the opening of the potassium channels is inversely proportional to I_{Kup}. If V_{mem} now drops below V_{thres}, the output of the first inverter V_{inv1} will become high, cutting off the current I_{Na}. Furthermore, the second inverter will then allow C_K to be discharged by the current I_{Kdown}. If I_{Kdown} is small, the voltage on C_K will decrease only slowly, and as long as this voltage stays high enough to allow I_K to discharge the membrane, it will be impossible to stimulate the neuron if I_{ex} is smaller than I_K. Therefore I_{Kdown} can be said to control the 'refractory period' of the neuron.

Finally I_{B1} and I_{B2} are two bias currents needed to limit the power consumption of the circuit; they do not influence the spiking behaviour of the neuron in any major way.

All current sources in Figure 5.1 are implemented with single transistors. These transistors of course will only behave like current sources when their drain-source voltage is larger than about 100mV, since they operate in weak inversion. If we ignore this limitation and assume that these transistors behave like current sources until their drain-source voltage becomes zero, we can adopt a piecewise linear approach. If furthermore we assume that $V_- = 0$, we can describe the circuit model with the following equations:

$$C_{mem}\frac{dV_{mem}(t)}{dt} = i(t) - I_{leak}, \qquad V_{mem}(t) < V_{thres} \qquad (37)$$

If $V_{mem}(t) = V_{thres}$: set $t_s = t$;

while $t_s < t < t_s + T_S$: set $V_{mem}(t) = V_+$;

while $t_s + T_s < t < t_s + T_S + T_R$: set $V_{mem}(t) = 0$, $\qquad (38)$

where $i(t)$ is the input current, C_{mem} the membrane capacitance, I_{leak} the leakage current, $V_{mem}(t)$ the membrane voltage, V_{thres} the threshold voltage, V_+ the power supply voltage, T_S the width of the spike, T_R the duration of the refractory period, and t_s is the time t when $V_{mem}(t)$ reaches V_{thres}. Equation (37) describes the leaky-integration of the input current on the membrane capacitance. In this circuit the leak comes not from a conductance in parallel with the membrane capacitance, but from a leakage current. In the piecewise linear approach, I_{leak} will flow as soon as $V_{mem} > V_-$. The membrane potential will decrease until it reaches its resting value as long as $i(t) < I_{leak}$. Only when $i(t) > I_{leak}$ will the membrane potential increase.

Equation (38) describes the reset operation. Whenever V_{mem} reaches the threshold V_{thres}, V_{mem} is drawn quickly to the positive power supply V_+, and after a delay T_S it is reset to its resting value, where it will stay for a duration T_R.

With a constant input current i ($i > I_{leak}$), the neuron will reach the spiking threshold in a constant time T_I after it leaves its refractory period. Integrating (37) and solving for $V_{mem} = V_{thres}$ yields:

$$T_1 = C_{mem} \frac{V_\theta}{i - I_{leak}} \tag{39}$$

The spike frequency is given by:

$$f_{spike} = \frac{1}{T_I + T_S + T_R} \tag{40}$$

An example of this spiking for a constant input current is given in Figure 5.2.
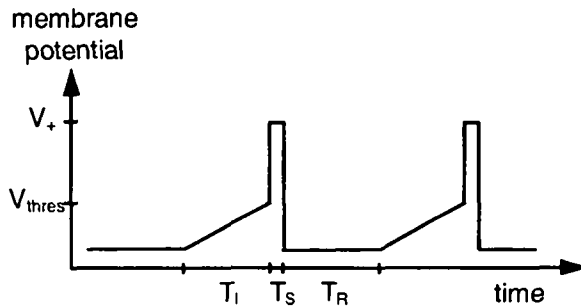

Figure 5.2 Typical evolution of the membrane potential in the neuron circuit for a constant input current.

The main difference between the (simplified) circuit model and the modified LIF neuron presented in the previous chapter (see also Tuckwell, 1988) is the fact that the circuit model has a leakage current and the LIF model a leakage conductance. In the neuron circuit, the membrane potential will stay at the resting potential when the stimulation current is smaller than the leakage current. When the stimulation current is larger than the leakage current, the membrane potential will always reach the spiking threshold after a time T_I directly proportional to $i - I_{leak}$. The leakage current thus not only functions as a leak which returns the membrane potential to its resting value, but also as a stimulation threshold. A stimulation current below this threshold will not cause spikes.

This is not that much different from the role of the leakage conductance g_L in the LIF neuron. In this model the threshold voltage V_θ will never be reached

when $i/g_L < V_\theta$, so that we can define a stimulation threshold equal to $g_L V_\theta$ below which the input current will never generate spikes.

5.3 Modelling cochlear nucleus neurons

5.3.1 Background

To test the circuit I have tried to model several different types of neurons in the cochlear nucleus, using different bias voltages and currents. The obvious reason to start modelling the auditory pathway with the neurons in the Cochlear Nucleus (CN) is that it is the first site of innervation of the auditory nerve. Furthermore, it is also the best-studied part of the auditory pathway. While many questions remain unanswered concerning the CN, we now know enough to begin the process of developing VLSI models of CN functioning. A surge of recent anatomical and physiological studies has revealed an unexpectedly complex structure at this early stage in the signal processing sequence. A recent collection of papers on the cochlear nucleus, edited by Merchan, et al., (1993) gives an authoritative account of current knowledge in this area.
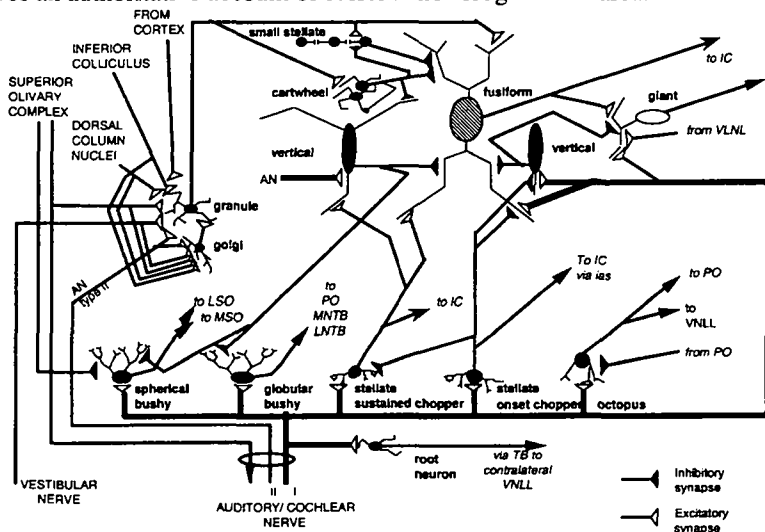

Figure 5.3 Different cell types in the cochlear nucleus and their innervation. Abbreviations: ias intermediate acoustic stria, AN auditory nerve, IC inferior colliculus, MSO/LSO medial/lateral superior olive, MNTB/LNTB medial/ lateral nucleus of the trapezoid body, PO periolivary nuclei, VNLL ventral nucleus of the lateral lemniscus. Adapted from Meddis and Hewitt (1993).

It is clear that a considerable amount of analysis takes place as soon as the signal enters the brain. This is achieved using a parallel processing strategy.

(Figure 5.3) When the Auditory Nerve (AN) enters the CN it immediately divides into two branches and later subdivides further. Each AN fibre makes contact with a large number of morphologically different cell types, each of which processes the signal in a different way. For example the AN fibres are known to innervate spherical bushy, globular bushy, stellate, octopus, vertical, giant and fusiform cells (Figure 5.3). Each of these cells reacts differently to a given acoustic stimulus, emphasising (or filtering out) particular features of the input. The various cells also project out of the nucleus to a range of different destinations. Even within a given morphological cell type, there are indications of further variation in the signal processing operations carried out. This is particularly true of the stellate cell class which appears to encompass a number of response types, including both excitatory and inhibitory ones.

Progress in modelling auditory signal processing has been greatly assisted by the observation that the variation in the responses of individual types of neurons can be understood in terms of two critical features of the cells: firstly, the input connectivity of the cell and secondly, the membrane characteristics of the cell. By input connectivity we mean the number of AN fibres contacting the cell, their origin along the cochlea and whether these inputs contact the cell soma itself or the cell dendrites. Synaptic events impinging on the soma give rise to rapid and strong membrane voltage changes at the axon hillock while contacts on the dendrites only give rise to slow and weak changes. By membrane characteristics we refer loosely to the response of the cell to depolarising and hyperpolarising (sub-threshold) currents. For example, Oertel (1983) has shown that stellate cells have linear voltage-current relationships (ohmic) while bushy cells show a saturating voltage-current relationship, i.e., the membrane voltage increases in response to an increase of current up to a certain voltage, after which the voltage tends to stay constant. This saturation is caused by an increase in the membrane's conductance resulting from the opening of fast potassium ion channels in response to depolarisation. This also results in a rapid membrane time constant when the cell is excited. Furthermore, the membrane characteristics are not only influenced by the types of ion channels in the membrane, but also by the density of these channels and the morphology of the cell. A larger cell will have a larger membrane capacitance and room for more leakage channels. The cell will then need more synaptic inputs, or stronger synaptic inputs in order to reach threshold. Differences in these key respects can result in substantial differences in the response of the cell to stimuli.

Almost all of the simulation work so far has used conventional computer programming techniques (see for instance Hewitt, et al., 1992 or Merchan, et al., 1993). This has proved satisfactory particularly when modelling the response of a single cell to a short burst of acoustical energy. However, the computer runs are typically very time-consuming and it is increasingly difficult

to envisage the simulation of large assemblies of such neurons without recourse to the real-time parallel processing capabilities of VLSI implementations.

The studies using computer models showed that the known physiological observations of single cells in the auditory brainstem can be successfully replicated using relatively simple models. This is true whether we use very detailed Hodgkin/Huxley type cell representations (Hodgkin and Huxley, 1953) or more schematic accounts such as those proposed by MacGregor (1987). It appears that good analogue models can be built if, as a minimum, the known connectivity is replicated and the cell membrane characteristics of the cells are reflected in the hardware design. The input to most of the cells in the ventral cochlear nucleus is dominated by direct connections from the auditory nerve, and therefore offers the possibility to test the influence of the circuit settings on the membrane characteristics, and to compare them with real neurons.

5.3.2 Methods

In this section we will test the functionality of the neuron circuit by comparing the response of the chip in different settings with the response of Ventral Cochlear Nucleus (VCN) neurons to tone bursts. VCN neurons are often characterised by their Post Stimulus Time Histogram (PSTH) — a histogram of the neuron's output spikes as result of repeated presentations of a stimulus, in this case a pure tone at a frequency that best stimulates the cell (for instance Rhode and Greenberg, 1992).

In our test, the tone burst is passed through the silicon cochlea, although the only effect of the cochlear filtering is to change the amplitude of the input signal somewhat, when measured at the output of the filter with a best frequency close to 5 kHz. The output of the cochlea is then passed through the inner hair cell circuit in order to half-wave rectify and low-pass filter the signal (see section 3.3). The IHC circuit also performs the temporal adaptation as shown in Figure 3.20.

The spiking neurons' PSTHs were created from the summed response of the 32 neurons on the chip to 20 presentations of the tone burst. If summation over multiple neurons is considered equivalent to summation over multiple stimulations, this can be thought of as the result of 640 stimulations. Statistically however both types of summation are not equivalent, although they both introduce variation in the neuron's response through device mismatch and noise respectively. Alternatively, we may think of both the neurophysiological and electrical PSTH as the result of repeated stimulations of a group of neurons. The neurophysiological PSTH is always measured on a single neuron, but the summed PSTH of a group of single neurons will have the same appearance as the PSTH of a single neuron in this group, if the neurons are almost identical.

5.3.3 Results

We will now briefly discuss the different types of neuron in the ventral cochlear nucleus of the cat, and compare their PSTHs with the PSTHs of the neuron chip in different settings. Most neurophysiological and anatomical information is taken from Popper and Fay (1992) and Webster, et al. (1992). Please refer to Figure 4.1 in chapter 4 to situate the different brainstem nuclei of the auditory pathway that will be mentioned in the following paragraphs.
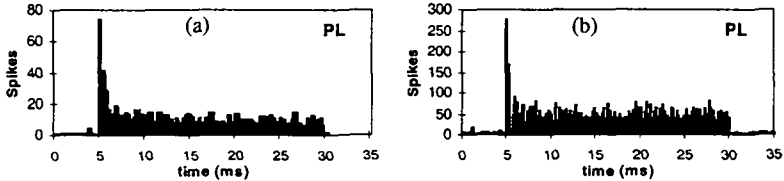

Figure 5.4 Primarylike PSTH. (a) Physiological response measured in the VCN of the Cat. (Based on data from Rhode and Greenberg (1992)). (b) Measured response of the chip.

Primarylike

The first type of response that I've tried to model is the Primarylike response. The Primarylike PSTH is the typical response of a spherical bushy cell, mainly found in the AnteroVentral Cochlear Nucleus (AVCN). This cell receives input from only one auditory nerve fibre through a large synaptic contact on its soma, called the endbulb of Held. This cell outputs one spike for every input spike, so it comes as no surprise that its PSTH is very similar to the PSTH of an auditory nerve fibre (compare Figure 5.4 with Figure 2.14 in chapter 2). The Primarylike cell is thus the least interesting to replicate, but is also traditionally the first VCN neuron presented in neurophysiological textbooks. In keeping with this custom it is also the first one to be presented in this section. It also shows that we can use this circuit to convert the output of the IHC circuit into spikes similar to those on the auditory nerve, should auditory nerve spikes be needed for a certain type of processing.

The spherical bushy cells also receive inhibitory inputs on their dendrites and might be implicated in some form of spectral sharpening. The Primarylike cells project almost exclusively to the Lateral Superior Olive (LSO) and the Medial Superior Olive (MSO), which are both part of the Superior Olivary Complex (SOC). The LSO and MSO are the brain nuclei that underlie sound localisation based on interaural intensity difference and time difference.

The chip's response was obtained by giving the neurons a medium-level membrane leakage current (I_{leak}) to avoid long term integration of the input signal and a short refractory period, so that the neurons are ready to fire when the next input arrives.
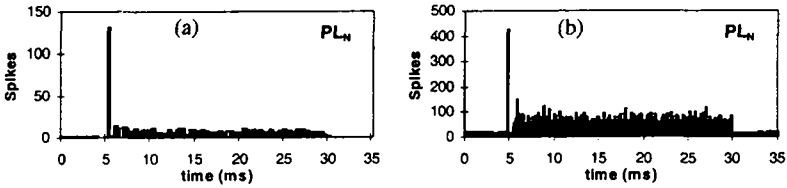
Figure 5.5 Primarylike with Notch PSTH. (a) Physiological response measured in the VCN of the Cat. (Based on data from Rhode and Greenberg (1992)).(b) Measured response of the chip.

Primarylike with Notch

A second type of response is the Primarylike With Notch response. Its PSTH is very similar to the Primarylike PSTH and is the typical response of a globular bushy cell, which is also mainly found in the anteroventral cochlear nucleus (AVCN). The globular bushy cell receives synaptic contacts from a few auditory nerve fibres that originate in the same region of the cochlea. Therefore, its PSTH is still similar to the PSTH of an auditory nerve fibre. However, the probability of the cell firing at the onset of the stimulus is almost one, because at least some of its input fibres will carry a spike at onset, due to the higher auditory nerve spiking probability at the onset of the stimulus (see Figure 2.14). If the cell always spikes at the onset of the stimulus, it will always be in its refractory period just afterwards. This creates the notch in the PSTH. The globular bushy cell thus enhances the onset of a stimulus. Furthermore, the cell typically needs several simultaneous auditory nerve spikes in order to generate an output spike. This means that the cell only responds to correlated action potentials of several fibres and thereby suppresses random activity present on the auditory nerve, particularly during silence. The globular bushy cell therefore may be said to improve the signal to noise ratio.

The globular bushy cells project to neurons in the Medial Nucleus of the Trapezoid Body (MNTB), which in turn project with inhibitory connections to neurons in the contralateral lateral superior olive. These LSO neurons also receive ipsilateral contacts from the spherical bushy cells. This neural circuit is probably the substrate for auditory localisation based on interaural intensity differences. The globular bushy cells also project to the Lateral Nucleus of the Trapezoid Body (LNTB) and to other PeriOlivary nuclei (PO). Both the LNTB and several other PO nuclei have efferent projections to the cochlear nucleus and the cochlea itself. Therefore, they are thought to be part of one or more feedback loops that modify the gain of the cochlea and that inhibit several neurons in the cochlear nucleus. Several of these periolivary nuclei also have reciprocal connections with other nuclei in the superior olivary complex and also with the Inferior Colliculus (IC), which is the next stage in the auditory

pathway towards the auditory cortex. The periolivary nuclei thus seem to regulate the operating point of many different neurons in the auditory pathway.

The Primarylike With Notch response is obtained from the chip by using a relatively high I_{leak} to avoid long term integration of the input signal, a threshold voltage that assures that the neuron will always fire at the onset of the stimulus, and a refractory period that matches the duration of the notch.
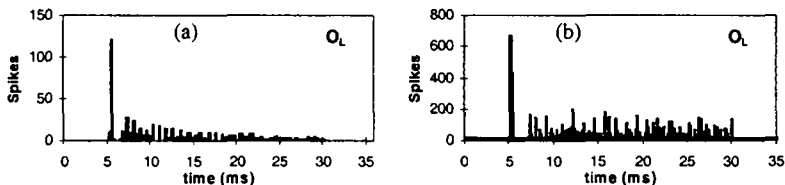

Figure 5.6 Onset Locker PSTH. (a) Physiological response measured in the VCN of the Cat. (Based on data from Rhode and Greenberg (1992)).(b) Measured response of the chip.

Onset Locker

A third response type is the Onset Locker response. Its PSTH is similar to the Primarylike With Notch PSTH, but with a lower and less constant sustained rate. This PSTH is the typical response of the large octopus cells found in the PosteroVentral Cochlear Nucleus (PVCN). It is obtained by the same high probability of spiking at the onset, which creates the onset peak and the notch in the PSTH. Having a higher leakage current than the globular bushy cell, and more synaptic inputs, it needs even more synchronous action potentials in order to generate a spike. Therefore, its signal-to-noise ratio will be even higher than the signal-to-noise ratio of the Primarylike With Notch response.

The onset spikes of the octopus cells have a very short latency, i.e., the cell reacts very quickly to the start of a stimulation, sometimes within 100 μs. Octopus cells project to some of the periolivary nuclei, which in turn inhibit the Outer Hair Cells (OHC) in the cochlea and also inhibit other cochlear nucleus neurons. They are therefore probably implicated in some fast feedback loops. They also project to the Ventral Nucleus of the Lateral Lemniscus (VNLL), which is known to be part of the neural circuit that mediates the fast acoustic startle response. Some neurons of the VNLL also have inhibitory projections to the Inferior Colliculus. Octopus cells might thus provide fast inhibition within the Inferior Colliculus. In general, octopus cells are probably part of the neural circuitry implicated in the detection of sound and the initiation of some reflexes.
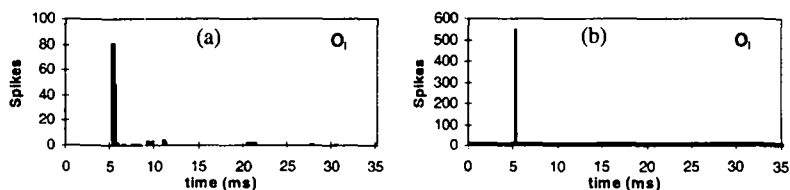
Figure 5.7 Onset Ideal PSTH. (a) Physiological response measured in the VCN of the Cat. (Based on data from Rhode and Greenberg (1992)).(b) Measured response of the chip.

Onset Ideal

A fourth type of response is the Onset Ideal response, which is a far less common response of the octopus cells. Its PSTH only shows an initial peak followed by hardly any activity. This could be because the input is only large enough to provoke firing at the onset of the stimulus. Alternatively, the cell might receive direct excitatory contacts from the AN and delayed inhibitory contacts. However, this inhibition has been shown to be absent in physiological experiments. The few cells that have been measured stay depolarised after the first spike, i.e., once the cell has fired, its membrane potential stays above the spiking threshold and the cell can not generate another spike as long as the input activity persists.

The Onset Locker and Onset Ideal responses are obtained from the chip by progressively increasing the leakage current from the setting that produces the Primarylike With Notch PSTH. Alternatively, the Onset Ideal response might be obtained by having an I_K only slightly larger than I_{Na}, so that the cell cannot reset as long as there is an excitatory input current.

Onset Chopper

A fifth type of response, the Onset Chopper response, is the response of a different neuron, the large, multipolar stellate cell. It is mainly found in the PVCN, but is also relatively frequent in the AVCN. This type of stellate cell has its soma largely covered by synaptic contacts, and also has a very short onset latency. The post-synaptic potential created by a spike arriving at a synapse on the cell soma will be less low-pass filtered than spikes arriving at synapses on the dendrites; each spike thus received will create a relatively abrupt membrane voltage variation. The cell receives many synapses from auditory nerve fibres and requires many inputs to fire. However, it has a smaller leakage current than the other two onset cells, so that the spikes are integrated over a longer time. This neuron is also thought to produce the Transient Chopper response, which is like the Sustained Chopper response (Figure 5.9), but with a less regular form

of chopping. This would also be caused by the many somatic contacts of the cell, generating many abrupt variations of the membrane potential.
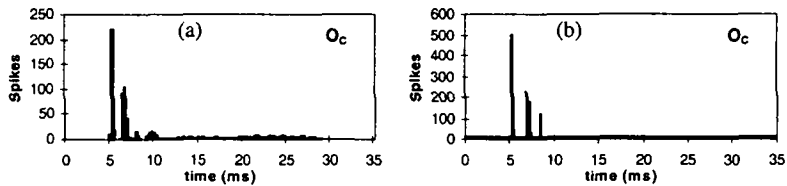

Figure 5.8 Onset Chopper PSTH. (a) Physiological response measured in the VCN of the Cat. (Based on data from Rhode and Greenberg (1992)).(b) Measured response of the chip.

Chopping is the standard behaviour of a spiking neuron that completely resets after spiking. After an action potential is generated, such a cell cannot be activated during its refractory period. Once its refractory period is over, the cell starts to integrate input spikes over a time that is controlled by the leakage conductance of this leaky integrator. This yields fairly regular firing in response to an input signal whose mean spike rate is constant, thus generating the typical chopping behaviour. If the sustained input level (from the auditory nerve) is too low to reach the firing threshold, the cell will only spike during the higher transient stimulation from the auditory nerve after a stimulus onset.

The Onset Chopper cell projects directly to the inferior colliculus and is therefore part of the shortest and fastest pathway from the auditory periphery to the forebrain. It also projects to several other neurons in the cochlear nucleus and to the superior olivary complex and the lateral lemniscus. The mean firing rate of the Onset Chopper increases with the intensity of the input signal and only saturates after an 80dB intensity increase of the input signal from the spiking threshold. Because of this very large dynamic range, this neuron is therefore thought to be part of the neural substrate that codes the intensity of sound. However, its output is inhibitory, so it only can serve to inhibit other neurons, for example to keep them at an operating point within their dynamic range.

This behaviour is obtained from the chip by having a relatively low leakage current and a high threshold voltage. Furthermore, an input signal with a high onset to sustained ratio is used, corresponding to the somatic contacts which perform less low-pass filtering than the dendritic contacts of the Sustained Chopper cells.
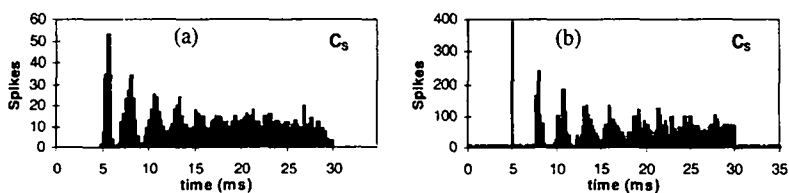
Figure 5.9 Sustained Chopper PSTH. (a) Physiological response measured in the VCN of the Cat. (Based on data from Rhode and Greenberg (1992)).(b) Measured response of the chip.

Sustained Chopper

Finally, the Sustained Chopper response is typical of another type of stellate cell. This stellate cell hardly has any synapses on its soma and thus receives most of its inputs on its dendrites. Because the dendrites low-pass filter the incoming spikes, the membrane potential will rise smoothly. This explains the regularity of this cells chopping response, shown in Figure 5.9. Regularity is however diminished due to the integration of noise over time, which adds a small random component to the interval between two successive spikes. The more spikes there have been between the current spike and the onset spike, the more uncertain its timing becomes. This effect is clearly visible for the last 10ms of the PSTH in Figure 5.9.

The Sustained Chopper cells have a dynamic range of only 30dB, much smaller than that of the Onset Chopper cells. They receive inhibitory connections from the Onset Chopper cells, which might keep the cell's operating point within its dynamic range. The Sustained Chopper cell has excitatory projections to the Inferior Colliculus and also to several neurons in the Dorsal Cochlear Nucleus. Together with the Onset Chopper cells they might be used to code intensity. Furthermore, they lock to the modulation frequency of amplitude modulated sound if the modulation frequency is well below 1kHz. Therefore, they are probably used in the extraction of the AM frequency and pitch of a speech signal (Hewitt, et al., 1992; Hewitt and Meddis, 1994). Several Sustained Chopper cells with the same best modulation frequency may project to the same coincidence-detecting neuron in the Inferior Colliculus, which is thus activated if a certain modulation frequency is present in the input signal. One of the system examples, which I will present in chapter 7, is based on this effect.

5.3.4 Discussion

Although the proposed neuron model is very simplistic, this neuron model already allows us to simulate the spiking behaviours characteristic of different neuron types by changing its biases. As shown in the previous section, Post

Stimulus Time Histograms closely resembling the PSTHs of ventral cochlear nucleus neurons can be obtained with the circuit. An advantage of the analogue VLSI implementation is that one may change the bias voltages and see the neuron model react in real time. This largely simplifies the task of determining the settings to use. Because of its simplicity, the neuron model is small; hundreds of neurons can be put on a reasonably sized chip. This makes it possible to implement and study the neural architectures of the auditory pathway and determine their utility as auditory signal processors.

Measuring the PSTH of a neuron, however, is not the only way to characterise a neuron and certainly doesn't characterise the neuron completely. One would have to measure spike rate versus frequency, spike rate versus intensity, with and without masking by another tone or noise; one would also have to generate the interspike interval histogram. The latter can be determined from the current set-up by measuring only a single neuron of the 32 on the chip. The measurement of the rate-intensity curve needs a model of the AN input which with the right output activity for a given input tone intensity. As I have shown in chapter 3, the time constant of the current IHC circuit depends on signal level, which makes it impossible to use this circuit in rate-intensity measurements without having to change the settings for the IHC circuit for every stimulus intensity. For the rate-frequency curve measurement, the IHC model needs to be connected to our electronic model of the cochlea, and the proper IHCs need to be connected to the neuron chip to model its frequency response.

Having an electronic model of a neuron is interesting in itself, but most of the neural processing in the brain is carried out by interactions between neurons. In the previous chapter, we have seen that (most) neurons communicate by sending spikes over axons which make synapses with the cell bodies and dendrites of other neurons. In this section, we will look in more detail at these different aspects of neural interaction.

5.4 Neural interactions

5.4.1 Spikes

The fact that the PSTHs of the circuit can be made to look similar to the neural PSTHs, even when the input of the neuron circuit is a continuous analogue signal (the output of the IHC circuit) instead of spikes, highlights a cardinal question in the research domain of neuroscience: "Why does the brain use spikes?"

There is an obvious, albeit not very satisfactory, answer to this question. It is clear that spikes are used to transmit information over relatively long distances in the brain and the entire nervous system. It would be impossible to

communicate analogue signals over these distances because of the internal resistance and membrane capacitance and leakage of the neuron. We might imagine points along the axon where the signal will be regenerated. However, if the waveform codes the information, we have to regenerate the exact waveform of the signal, which is an impossible task. If the waveform does not code the information, then there will be no need to communicate the complete analogue waveform, and we could use spikes for instance. It is generally accepted that the exact form of a spike is not important, but that the spike signals an event, and that most information is coded by the timing of the spike, the average rate of spikes on the axon, and/or the position of the neuron which originates the spike.

Spikes are indeed regenerated along the axon at the nodes of Ranvier. A spike arriving at such a node will depolarise the local membrane potential, and trigger the spike generation process. As we have seen in the previous chapter, the spike generation process includes a refractory period, and this ensures that the spike only travels in one direction along the axon. Actually, the regenerated spike will travel in both directions along the axon to its neighbouring nodes of Ranvier, but since the node closer to the axon hillock has just spiked before, it will be in its refractory period. This means that it impossible to trigger the spike regeneration process at this node and that the voltage-dependent potassium channels are still open, which shunts further diffusion of the spike in this direction. The node of Ranvier farther away from the axon hillock will not be in its refractory period, and here the regeneration process will be activated. Without a refractory period in the spiking mechanism each node of Ranvier would stimulate the others when it spikes, and spiking would never stop.

Just as the spiking itself, the refractory period is thus needed for communications between neurons. However, as we have seen in section 5.3.3 when we discussed the chopping behaviour of some neurons, the refractory period is also an important factor which determines the maximum chopping rate of the neuron. In the next chapter, I will show that this can be used to tune the chopper and make it sensitive to particular amplitude modulation frequencies in its input signal. This example shows that although the refractory period is necessary to transmit spikes over an axon, it also serves an important function in the signal processing performed in the brain. This brings us to the actual question about spikes: "Are spikes just needed for communication in the brain, or do they also play a major role in neural signal processing?" And, more importantly to us engineers: "Does signal processing with spikes offer important advantages over other types of signal processing?"

The answer to the first question is certainly that spikes do play an important role in the actual signal processing, for the simple reason that the brain evolved as a system which uses spikes as its information carrier. However it is still

largely unclear what the signal processing operations are that the brain performs in treating the auditory sensory input with spikes. What is more, the way in which information is coded with spikes is still an issue of debate, and as long as we don't understand how the information is coded, then how can we hope to understand the operations that the brain performs on the information?

Information can be coded by an average rate of spikes, averaged either in time, or over a group of similar neurons, or both. In this case the precise timing of the individual spikes has no importance. On the other end of the scale, information can be coded by the arrival time of an individual spike; all the information is in the timing. The brain, however, does not have to choose between one code or the other and uses both, sometimes even at the same time. Furthermore, the brain will have to use group coding, be it a temporal or average rate code, because single neurons have too low maximum spike rates and are too prone to error, noise or cell death. For this reason all computation in the brain will have to be collective, and for the same reason collective computation is an interesting option for analogue VLSI. Especially in the lower centres of the auditory pathway, timing seems to be important, but it costs effort to keep the timing between different events in the different stages in the auditory pathway, and I believe that at the higher stages of processing the brain uses codes which are more like average rates.

Relative timing of spikes hints at signal processing operations like coincidence detection and correlation of different signals, a multiplicative operation. On the other hand, average rates represent analogue variables, and hint of addition and subtraction of signals using excitatory or inhibitory synapses. It has been recently shown by Maass (1996, 1997) that it is theoretically possible to build a Turing machine from multilayer perceptrons using temporal coding with spikes, and leaky-integrate-and-fire neurons. Most of the standard neural network theory to date however has been based on the assumption that the information is coded by spike rates. These mathematical constructions tell us that in principle all computational operations are possible with neurons using either temporal coding or average rate coding. They do not tell us, however, which operations can be implemented efficiently and which cannot.

Another approach is to look at operations we can perform particularly effectively with spikes. Three such operations are discussed below; all are believed to occur in the brain.

Perhaps the simplest such operation is coincidence detection: using spikes, only a digital AND operator is needed to implement it. The brain, however, does not use an AND operator, but a spiking neuron to implement coincidence detection. This results in fuzzy coincidence detection; because an incoming spike creates a change in a neuron's membrane potential lasting longer than the

spike itself, a second spike can add to the influence of the first spike. Although membrane depolarisation is greatest when the spikes coincide, two not-quite-coincident spikes are still more likely to generate an action potential than a single, isolated spike. The temporal sensitivity of this process, controlled by the leakage conductance and membrane capacitance of the neuron, will be different for different cell types.

Applying fuzzy coincidence detection to spike trains results in synchronicity detection; a coincidence detecting neuron will fire most when two spike trains are synchronous; its response will decrease as the synchrony between the two spike trains decreases.

As third operator, consider cross-correlation of two signals: synchronicity must be detected between one signal and delayed versions of the other. A delay line is simpler to implement for spikes than for analogue signals, because no information is contained in the form of the spike; we can just regenerate each spike after a certain delay. In the case of an analogue waveform, we have to somehow reproduce the entire waveform after a certain delay.

In summary, one advantage of spike coding seems to be the fact that it is simple to implement synchronicity detection, which provides a powerful way of comparing two temporal signals. The ease with which coincidences can be detected also allows the brain to use synchronicity as a way of coding information. It has been shown for instance that many different types of neurons in the cochlear nucleus are synchronised by amplitude modulation at a particular modulation frequency; in absence of this modulation component these neurons automatically desynchronise (Frisina, 1990; Rhode and Greenberg, 1994). The synchronisation of a certain group of neurons is thus a way to code the presence of a particular feature, the amplitude modulation frequency in this case. In chapter 7 we will see an example where the synchronisation of sustained chopper neurons from the ventral cochlear nucleus and coincidence detection performed by inferior colliculus neurons are used to extract this amplitude modulation frequency.

5.4.2 *Inter-chip communication with spikes*

Although it has nothing to do with the spiking of neurons in the brain, there is a good engineering reason to adopt spike coding of signals for inter-chip communication. Whenever we have a large number of outputs in parallel, as is the case when we have a large group of neurons on chip, we cannot simply connect each output to another chip with a dedicated wire. The traditional solution in this case is to scan serially through the output. This means that at each clock cycle the output of a different element is read, so each output is only read once every N clock cycles, where N is the total number of outputs. When N grows large, the clock frequency will also have to become large in order to

maintain enough temporal resolution on each output. In the case of a large array of neurons, we would be scanning all neurons, most of which would be inactive most of the time. It makes more sense just to transmit the fact that a certain neuron spiked at a certain time. This type of communication has been termed 'event driven communication' (Lazzaro, et al., 1993; Mortara, 1995).

Rather than providing a wire for each output, we can communicate spike events using a single data channel shared by all the neurons on a chip. Each neuron has a unique address; when a neuron spikes, its address is sent onto the channel. The address pulse can be very short if the address detector on the other end is fast enough; we can regenerate a spike with biological duration on the receiving end of the data channel. With spikes lasting only about 1 μs and a maximum spike rate limited to a few hundred Hertz (as it is in biological neurons), even the most active neuron will occupy the data channel less than 0.1% of the time.

In the most direct implementation of this idea (Mortara, 1995), every neuron may access the data channel at any time, which thus preserves perfectly the time structure of the output signals. The only problem with this communication is that it is possible for two or more neurons to access the data channel at the same time, resulting in an address collision. Using a special address coding scheme, for instance always the same numbers of ones and zeros in the address code of each neuron, we can detect these collisions and decide to ignore these addresses. This introduces noise in the communication process, but if collisions are rare, this noise level is very low.

One situation, in which the chance of having collisions is increased, is when information is coded by synchronous activity of a group of neurons. This situation is of particular interest to us and warrants special attention, since we have identified synchronicity as an important property in spike-based computation. In this case several neurons spike at the 'same' time to signal a salient feature. The 'sameness' in time of two spikes is however measured at a biological time scale, in the order of 1 ms. So two spikes about 100 μs apart are still considered synchronous on this scale, but on the time scale of the communication process they can be quite far apart, if for example each communication pulse only lasts 1 μs. After regenerating spikes of biological duration on the other end of the data channel, we can easily detect the synchronicity of the spikes on the biological time scale again. The problem is therefore not as bad as it looks, but still it increases the probability of collisions with respect to an activity of the neurons which is evenly distributed in time.

I have used this communication scheme for the neuron chip discussed in section 5.2. The circuit for the input decoding and output encoding is shown in Figure 5.10. The address of each of the 32 neurons on chip is coded with three ones and four zeros, yielding a seven-bit communication bus. Note that the

minimum number of bits needed to code for 32 addresses is five, so the
overhead created by this encoding (which allows collision detection) is only
two wires. At the input side of the array, an address-decoding block, stimulates
with a spike the neuron at a position corresponding to the address that has been
detected. The address-decoding block may be set either to stimulate none of the
neurons when an address collision is detected (col=1), or to stimulate all the
neurons with addresses that could have created this collision code (col=0).

in-bus col   sum out-bus
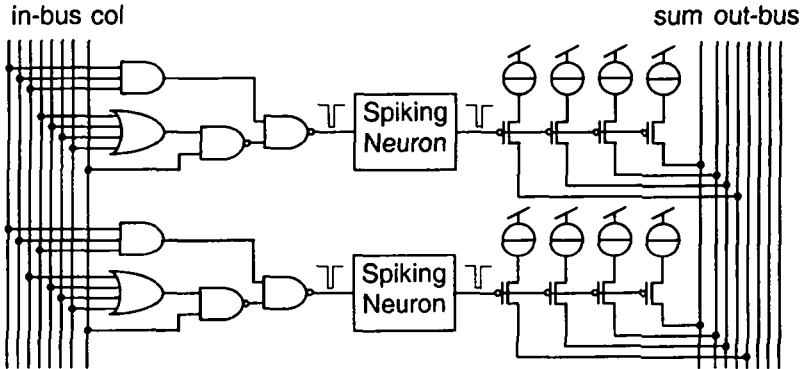
Figure 5.10 Input and output coding of the neuron chip

When a neuron spikes, apart from sending its address on the bus, it also sends
a current spike on a wire common to all neurons. This allows us to measure the
number of spikes by measuring the total current on the wire and dividing it by
the unit current used to send a single spike. This has been used in section 5.3.3
to create the PSTHs in that section.

No circuits were included to reduce the spike width at the output side or to
restore the biological spike width at the input side. With only 32 neurons on a
chip, collisions will not happen very often even with the wider spikes, unless
we are synchronising the neurons with the input signal. If that is the case, we
will be interested specifically in the synchronicity of the neurons on the chip;
we can use the single common output wire to detect this.

Note that the address is coded on the output bus by current pulses, but by
voltage pulses on the input bus. We have to use current pulses on the output bus
because all neurons will drive the same bus wires, and currents injected on a
single wire will just add up. Voltages cannot be used directly because if two
neurons drive the same wire, one trying to impose 0V and the other 5V, we will
have a conflict, the voltage on the wire will be undefined, and a lot of current
will flow through the output transistors of the two neurons. Using current pulses
on the output bus also allows combining the output buses of two chips into a
single bus, i.e., combining two seven-bit buses into a single seven-bit bus. In

this way neurons with the same address on the two chips will both send their pulses to a neuron which detects this address on a third chip.

For the input bus, however, voltages are preferable so that we can contact the inputs of each address decoding block with the same bus wires. Each chip will receive input addresses coded by current pulses and will convert these to voltage pulses. Furthermore, the output pin of the sending chip and the input pin of the receiving chip, plus the external wire, will add a substantial capacitance to the communication link. Therefore we would need a lot of current to change the voltage on the wire quickly. However, using a virtual ground at the input of the receiving chip to sense the current, we can keep the voltage on the wire constant. The circuit which does this for each input wire is shown in Figure 5.11a. The structure with the two PMOS transistors in the top left of the circuit, together with the current source I_{cc}, serves to keep the voltage on the input line constant while conveying the input current to an internal node. The input current is then copied by the current mirror, and this current is compared with the current I_{cmp} to determine if the line is at 1 or 0.
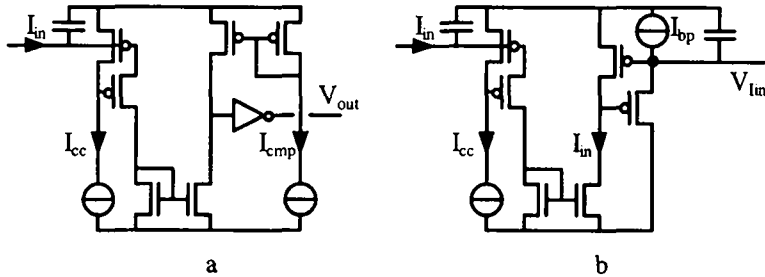

Figure 5.11 Current conveyors: a) with comparator, b) with line driver.

In some cases though, the input to the chip will not be in the form of pulsed addresses, but in the form of an input current, as in the measurements of section 5.3.3, where the input of the chip is the output current of the IHC circuit. Again, we can hold the voltage on the input line constant and use a current mirror with 32 outputs to create copies of the input current for each neuron. Creating 32 copies of the current means driving 33 gates which together have a considerable capacitance. To drive this capacitance, we can use the inverse of the current conveyor structure, as shown in the top right of Figure 5.11b. This avoids drawing the current needed to charge or discharge the capacitance from the input current. We can thus make more current available to drive the capacitance than the input current would allow, which will speed up the communication.

5.4.3 Synapses

In the previous chapter, we have seen that neurons communicate through synapses and that the result of a spike arriving at a synapse can be excitatory, inhibitory, or shunting inhibitory, depending on the neurotransmitter released by the pre-synaptic cell, and on the type of receptors in the post-synaptic cell. An excitatory input will increase the membrane potential; an inhibitory input will decrease the membrane potential, and a shunting inhibitory input will draw the membrane potential towards the resting potential. The effect of these synapses can be modelled with the circuit of Figure 5.12a. The thin lines represent spikes arriving at the different types of synapses. Note that to control the excitatory synapse, the spike has to be inverted with respect to the biological spikes. The figure also shows the part of the neuron circuit that models the passive membrane properties, i.e., the membrane capacitance, and the leakage current towards the resting potential.
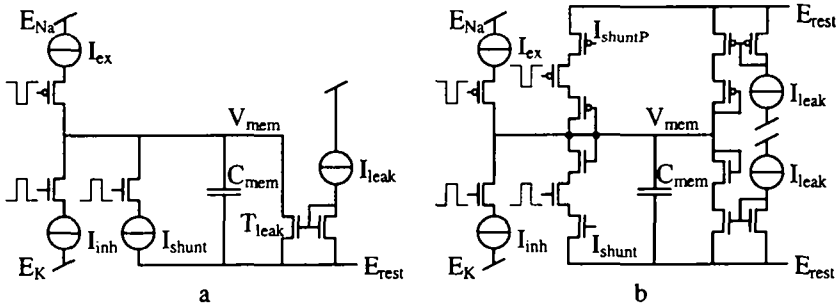

Figure 5.12 Circuit for three different types of synapses. a) basic circuit, b) corrected for hyperpolarisation.

In this circuit, an excitatory input will increase the membrane potential by an amount:

$$\Delta V_{mem} = \frac{T_S(I_{ex} - I_{leak})}{C_{mem}}, \tag{41}$$

where T_S is the duration of the spike. Both the inhibitory and the shunting inhibitory decrease the membrane potential by a similar amount. However, this is only true when V_{mem} is above E_{rest}, because the output current of transistor T_{leak} is only equal to I_{leak} when V_{mem} is at least about 100 mV larger than E_{rest}. When V_{mem} is equal to E_{rest}, the current through T_{leak} is obviously zero.

The implementation of Figure 5.12a is limited to $V_{mem} \geq E_{rest}$; when V_{mem} is below E_{rest} the current through T_{leak} will be inverted, effectively drawing the membrane potential up towards the resting potential, but the amount of current with which it will do this will now increase exponentially with decreasing V_{mem}.

Therefore, this current will get very large for a V_{mem} only a few hundred millivolts below E_{rest}. A simple way to attempt to correct this problem would be to use a diode-connected transistor in series with T_{leak} so that the current can only flow through this branch when V_{mem} is larger than E_{rest} and add a second branch using PMOS transistors to model the leakage current when V_{mem} is below E_{rest}. Furthermore, because the current source I_{shunt} will be implemented in a similar way as I_{leak}, we will have to do the same for the shunting inhibitory synapse. This then yields the circuit of Figure 5.12b.

The problem with this solution is that normally E_{rest} will be closer to E_K than to E_{Na}. The PMOS transistors that have their bulk at the E_{Na} and their source at E_{rest} will thus have their threshold voltage increased due to the body effect. This will mean that these branches can only conduct properly when V_{mem} is well below E_{rest}, which is not possible if E_{rest} is close to E_K. Furthermore, we cannot solve this by placing the PMOS transistors in well that is tied to E_{rest}, because this would forward-bias the drain-well junction of the diode-connected transistors when V_{mem} rises above E_{rest}. We could design a bias circuit that correctly biases the well of the PMOS transistors as a function of Vmem, but it is clear that having hyperpolarisation and inhibitory synapses complicates the circuit considerably.

The neuron chip presented in section 5.2 does not allow for hyperpolarisation. This also means that inhibition can only be in the form of shunting inhibition, because the inhibitory synapse can not draw the membrane potential to a value lower than its resting potential. The two possible synapses for this neuron are thus the excitatory and shunting inhibitory synapses shown in Figure 5.12a.
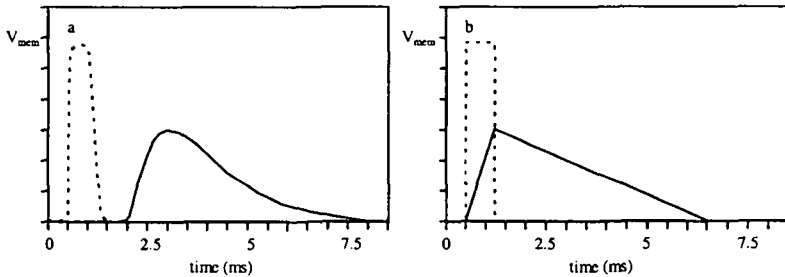

Figure 5.13 Schematic representation of an EPSP in a biological neuron (a) and in the electronic neuron (b). The dotted trace represents the timing of the pre-synaptic membrane depolarisation, but is not necessarily on the same voltage scale. The post-synaptic depolarisation is often much smaller than the pre-synaptic depolarisation.

In the biological neuron, the change in membrane potential is mediated by changes in the ion conductances and thus a voltage-dependent current injection, whereas in the electronic circuit it is caused by direct (voltage-independent)

current injection. This causes some differences in the shape of the Excitatory Post Synaptic Potential, as shown in Figure 5.13 a and b. However, in both cases the EPSP shows a steep rising flank and a shallow falling slope, the most important features of the EPSP shape. An excitatory input will have its maximum effect only at a certain time not too long after the arrival of the spike at the synapse and after this time the influence of the excitatory input on the membrane potential slowly decreases. This means that two spikes arriving at the same time at two excitatory synapses will yield a larger increase of the membrane potential than two spikes that arrive within a short time of each other, and two spikes arriving within a long time interval will have the same effect as just the second spike by itself.

Another difference between the biological synapse and the electronic synapse is that the chemical transmission at the synapse of the biological cell causes a synaptic delay, which is absent in the electronic circuit. Although this changes the absolute timing of the spikes, it does not alter the relative timing of the spike, at least not if the biological synaptic delay is considered constant. Of course, the relative timing between two different pathways is only retained if the number of neurons in each pathway is the same. Although an extra synapse in one pathway will not change the delay in the electronic circuit, in both the biological and the electronic case an extra neuron will add a considerable delay to this pathway. If we take care that the electronic neuron adds a delay in the pathway which is equal to the sum of the neural and synaptic delay in the biological case, then both systems are equivalent again, from at timing point of view.

5.4.4 Dendrites

Equation (41) shows that the synaptic strength can be controlled using I_{ex} or the spike width T_S. This means however that we need a different current source for each synapse, or a way to adapt the spike width locally at the synapse, since a neuron sends a spike of the same width to all the synapses it makes with other neurons. Furthermore, the output of the neuron will have to be physically connected with a wire to each synapse it makes, and these connections would use up most of the silicon area when each neuron makes multiple synapses with other neurons. As we have seen in the previous chapter, the effect of an excitatory input at a synapse on the membrane potential also depends on where the synapse is situated on the dendritic tree. Because of the internal resistance of the dendrites combined with the membrane leakage current, a synapse situated more apically on a dendrite will change the membrane potential at the axon hillock less than a synapse closer to the cell body, even if they both create the same local membrane depolarisation. The circuit of Figure 5.14 models this effect. (See also Elias, 1993 for electronic-dendrite models.)

At the top of Figure 5.14 we recognise a number of excitatory synapses, modelled as before by current sources with switches controlled by the input spikes. When a spike arrives at synapse 3 for instance, the current I_{syn3} will flow through the switch for the duration of the spike. Part of this current will flow directly through the conductance G_3; the output current I_{ex3} will be equal to this current. The rest of the current flows into the network to either side of the node which connects the different synapses laterally. If we assume that the network extends infinitely to the left and to the right, equal amounts of current will flow left and right. At the neighbouring nodes these currents will split again, partially continuing to flow through the lateral transistors, and the rest creating the output currents I_{ex2} and I_{ex4}. In this way, a current injected at a certain node in the network creates a symmetrical distribution of output currents; The largest output current occurs at the same node as the injection, and the amplitudes of the output currents decrease as the node gets farther away.
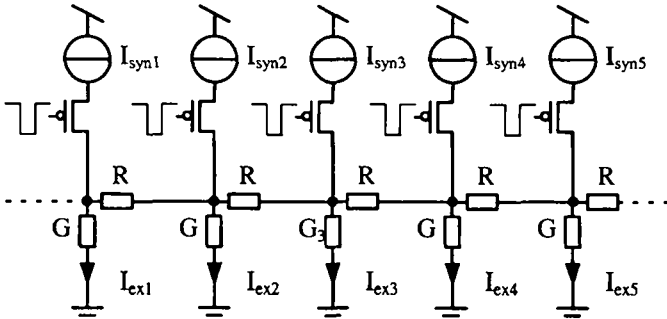

Figure 5.14 An electronic dendrite with resistors.

Mead (1989) has shown that in such a resistive network the voltage on node j as a function of the voltage at the input node i can be approximated by:

$$\frac{V_j}{V_i} \approx e^{\frac{-d_{ij}}{L}}, \qquad L = 1/\sqrt{RG}, \quad L > 1 \qquad (42)$$

where d_{ij} is the distance between nodes i and j measured as the number of intervening lateral resistors, R the value of the lateral resistance, and G the value of the vertical conductance. We can calculate the current distribution from this equation, since the current I_{exj} through the output conductance at node j is proportional to the voltage on node j and an input current I_{syni} at node i will create a certain node voltage V_i. We can therefore write:

$$\frac{I_{exj}}{I_{syni}} \approx \alpha e^{\frac{-d_{ij}}{L}}, \qquad L = 1/\sqrt{RG}, \quad L > 1 \qquad (43)$$

where α is a constant. We can calculate the value of α using the fact that the sum of all output currents has to be equal to the input current. If we assume that the resistive network is infinite, we can write:

$$\sum_{d=-\infty}^{\infty} \alpha e^{\frac{-d}{L}} = 1 \tag{44}$$

where d is the distance between the output node and the input node measured in lateral elements. Because the distribution of output currents is symmetrical about d=0 we can also write:

$$-\alpha + 2 \sum_{d=0}^{\infty} \alpha e^{\frac{-d}{L}} = 1 \tag{45}$$

where the '$-\alpha$' term corrects for the fact that we count the term at d=0 twice. The summation term is a standard geometrical series, so that we can solve for α:

$$\alpha = \frac{1 - e^{-1/L}}{1 + e^{-1/L}} \tag{46}$$

Hence, we can determine α as a function of L.

It has been shown (Vittoz, 1994) that such a network of resistors as in Figure 5.14 can be implemented with MOS transistors in weak inversion, as long as we are only interested in the input and output currents and not in the voltages on the nodes. This can be explained by the following analysis.

The current through a MOS transistor operated in weak inversion as a function of the voltages on its terminals referred to the local substrate is given by:

$$I_{DS} = I_S \, e^{\frac{V_G - V_{TO}}{n \, U_T}} \left(e^{\frac{-V_S}{U_T}} - e^{\frac{-V_D}{U_T}} \right) \tag{47}$$

with V_G the voltage at the gate, V_S the voltage at the source and V_D the voltage at the drain terminal respectively, all referred to the local substrate. V_{TO} is the threshold voltage of the MOS transistor, U_T is the thermal voltage and I_S is the specific current (defined in chapter 3, equation (2)).

The linear behaviour of the transistor in the current domain can be made visible by defining a pseudo-voltage V^* as:

$$V^* = - V_0 \, e^{\frac{-V}{U_T}} \tag{48}$$

where V_0 is and arbitrary scaling value. We can then rewrite equation (47):

$$I_{DS} = g^* (V_D^* - V_S^*)$$ (49)

with:

$$g^* = \frac{I_S}{V_0} \, e^{\frac{V_G - V_{T0}}{n \, U_T}}$$ (50)

This shows that with respect to the current, the MOS transistor in weak inversion behaves as a conductance of which the value can be controlled by its gate voltage. Furthermore, from (48) we can see that as soon as V_D becomes larger than a few U_T, V_D^* becomes almost zero and the drain terminal can thus be considered as a pseudo-ground. We can combine these pseudo-conductances in a network as long as we use the same definition of pseudo-voltage for all transistors, and the same V_0.

It is not a real constraint in our case to operate the transistors in weak inversion. If we want small neurons, the membrane capacitance C_{mem} in the neuron circuit of section 5.2 will have to be small (a few pico-Farad) since a capacitor consumes a large silicon area. Furthermore, the spike width T will be close to 1 ms to stay in the same range as the biological spikes. Equation (41) shows that I_{ex} and thus I_{syn} will have to be small if we want to limit ΔV. So the transistors in the dendritic tree can easily be operated in weak inversion.

Replacing the conductances G and R in Figure 5.14 with pseudo-conductances T_G and T_L, results in the circuit of Figure 5.15.
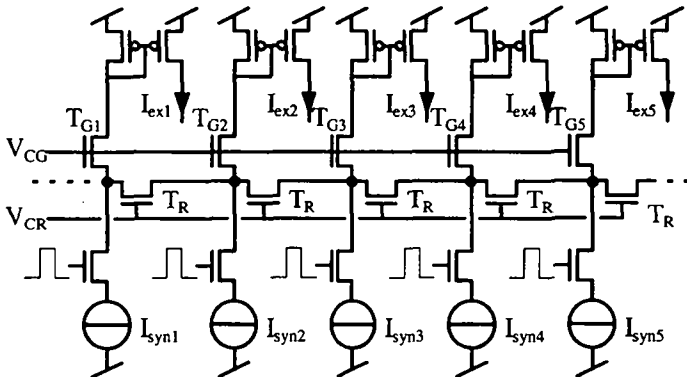

Figure 5.15 Electronic dendrite with pseudo-conductances.

Equation (43), still valid for this circuit, is now a function of the pseudo-conductances:

$$\frac{I_{exj}}{I_{syni}} \approx \alpha e^{\frac{-d_{ij}}{L}}, \quad L = \sqrt{(1/R)^*/G^*}, \quad \alpha = \frac{1 - e^{-1/L}}{1 + e^{-1/L}} \tag{51}$$

We can use equation (51) to express the pseudo-conductances G^* and $(1/R)^*$ as a function of the control voltages V_{CG} and V_{CR}. Assuming that T_R and T_G have the same geometry, we can write for L:

$$L = e^{\frac{V_{CG} - V_{CR}}{2 n U_T}} \tag{52}$$

Together with equation (51) this shows us how the strength of a synapse i decreases with increasing distance, when seen from node j. Now imagine that each output current I_{exj} charges the membrane capacitance of a different neuron. Equations (51) and (52) then describe how a given synapse i influences a neuron at position j. Because the diffusion network is completely linear, we can apply the superposition principle. Summing up the influence of all synapses on each neuron, we see that although all synapses influence all neurons, each input has maximum effect on the neuron it is closest to; its strength decreases with increasing distance. This structure can also be extended to a two-dimensional structure in which each input connects to a group of neurons in a 2-D map, and the strength of an input is maximal at the position of its synapse and falls off with distance.
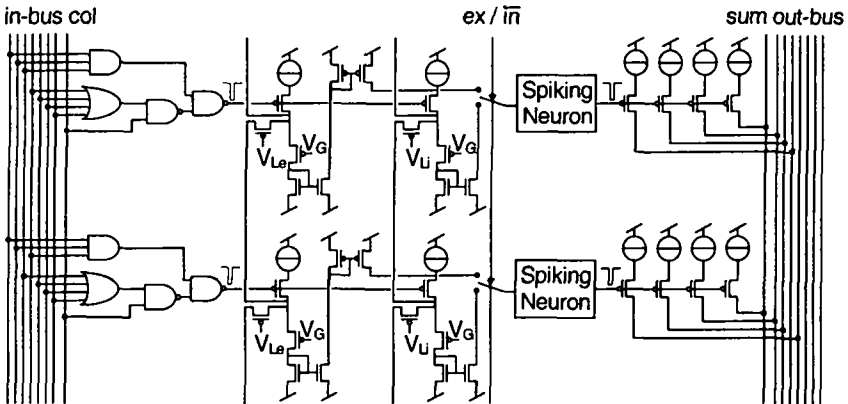

Figure 5.16 Input and output structure of the neuron chip, including two dendritic structures.

The dendritic structure gives us a nice way to connect the output of a neuron to a group of other neurons which are close to each other in an array, without the need to draw a connection wire from this neuron to all the neurons it influences. I have included two such structures at the inputs of the 32 neurons

on the neuron chip of section 5.2, one for the excitatory synapses, and one for the inhibitory synapses. This allows the excitatory and inhibitory inputs to have different diffusion length. Each address received from another chip will have one bit indicating if it is excitatory or inhibitory in addition to the seven bits which codes for position. Figure 5.16 shows the final structure.

The circuit of Figure 5.15 does not model the effect that a synapse more distal on the dendrite will generate a post synaptic potential that starts later and lasts longer in addition to being weaker, because the dendritic membrane capacitance has not been modelled in this circuit. We could model these temporal effects by adding a capacitor at each node in the circuit of Figure 5.14. However, in this case we are not only interested in the network currents anymore, but also in the node voltages at the nodes to which the capacitors are connected. If we want to do this for the circuit with the pseudo-conductances in Figure 5.15 we have to add a (pseudo) capacitor, which is compatible with pseudo-voltages in this circuit. Although this is certainly possible, it will take up space, and since I have not had the need for such a circuit, it has not been implemented.

5.4.5 Axons

In principle it is not necessary to model the regeneration of a spike in an axon at the nodes of Ranvier, because the resistance of the metal wires on chip is low enough to ensure proper transmission of a spike. However, a biological axon will also add a delay which increases with increasing length of the synapse. This delay may be used in certain neural computations; it is a lot easier to model than the dendritic delay because the form of the spike remains the same. In principle, we may use the same circuit as the neuron circuit in Figure 5.1 to model the nodes of Ranvier, since both have the same membrane properties. However, in the neuron circuit C_{mem} models the capacitance of the cell body, whereas for the node of Ranvier circuit we only have to model the capacitance of the node plus a section of the axon. We can thus use a smaller capacitance in this circuit. Moreover, since in the electrical circuit there is no coupling from the output of the circuit to its input, unidirectional transmission of the spike is guaranteed and we do not need to model the refractory period of the neuron in the node of Ranvier circuit. Modelling the refractory period in the neuron circuit of Figure 5.1 will be sufficient to ensure that there will be no spikes on the axon for a certain period after each spike. This means that the capacitor C_K can be reduced in size too, since it is only used for controlling the spike width, which is in general much shorter than the refractory period of the neuron. For the input of the node of Ranvier circuit we can use the same circuit as for the excitatory synapse, which will allow us to control the delay created by the circuit. Since this circuit only charges the capacitance during the input spike,

the delay created by the node of Ranvier circuit will not be larger than one spike width, because each input spike has to create an output spike. For this same reason, we do not need to model the leakage properties of the membrane, which simplifies the node of Ranvier circuit even more.
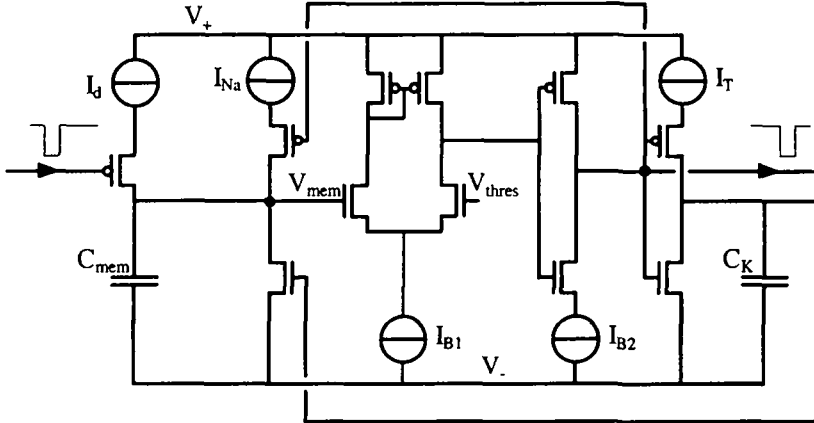

Figure 5.17 The node of Ranvier circuit.

5.4.6 Discussion

In this section we have looked at the different elements of neural interactions and presented the circuits for several of them. In contrast with previous sections concerning neural hardware we have presented no measurements in this section (5.4). It is of course trivial to characterise the synapse circuits, since they are nothing more than current sources with a switch in series. The dendritic circuit has been successfully used on the neuron chip in combination with both excitatory and inhibitory synapses, and also in a neural system for pitch detection which I will present in the next chapter. However, this circuit is embedded within these systems, and no direct measurements of its output currents are possible. The node of Ranvier circuit is the only circuit that has not actually been built, but its correct operation can be inferred from that of the neuron circuit with I_K and I_{Kdown} set to large values and I_{leak} set to zero.

5.5 Summary

In this chapter I have presented an electronic circuit modelling the spike generation process in the biological neuron. This neural circuit integrates charge on a capacitor which models the membrane capacitance of a nerve cell. When the voltage on the capacitor reaches a certain threshold value, a positive feedback cycle is activated, which quickly increases the capacitor voltage. After a short delay, a second feedback restores the resting voltage on the capacitor.

After the generation of a spike, during the refractory period, it is harder or impossible to create a new spike, due to the dynamics of this second feedback.

Comparison of the output of the circuit with the output of biological neurons in the ventral cochlear nucleus in response to tone bursts filtered by the cochlea and inner hair cells shows that the circuit is able to model the spiking behaviour of different types of neurons by changing its biases.

Neural computation obtains its power not from a single neuron, but from the interaction between a large number of neurons. Neurons typically communicate by sending spikes over axons which make synaptic contacts with the dendrites and cell bodies of other neurons. Circuits that model these interactions have also been presented in this chapter. They include the circuits for excitatory, inhibitory and shunting inhibitory synapses, a circuit which models the regeneration of spikes on the axon at the nodes of Ranvier, which can be used as a delay line, and a circuit which models the reduction of input strength with the distance of the synapse to the cell body on the dendrite of the cell.

5.6 References for chapter 5

Elias, J. (1993) "Artificial dendritic trees," *Neural Computation*, Vol. 5., pp. 648-663.

Hewitt, M.J., Meddis, R., and Shackleton, T.M. (1992) "A computer model of a cochlear-nucleus stellate cell: Responses to amplitude-modulated and pure-tone stimuli." *Journal of the Acoustical Society of America*, Vol. 91, pp. 2096-2109.

Hewitt, M.J. and Meddis, R. (1994) "A computer model of amplitude-modulation sensivity of single units in the inferior colliculus." *Journal of the Acoustical Society of America*, Vol. 95, pp. 1-15.

Hodgkin, A.L. and Huxley, A.F. (1953) "A quantative description of membrane current and its application to conduction and excitation in nerve." *Journal of Physiology*, Vol. 117, pp. 500-544.

Lazzaro, J., Wawrzynek, J., Mahowald, M., Sivilotti, M., and Gillespie, D. (1993) "Silicon auditory processors as computer peripherals," *IEEE Journal of Solid State Circuits*, Vol. 26, pp. 523-528.

Maass, W. (1996) "Lower bounds for the computational power of networks of spiking neurons," *Neural Computation*, Vol. 8, pp. 1-40.

Maass, W. (1997) "Fast sigmoidal networks via spiking neurons," *Neural Computation*, Vol. 9, pp. 279-304.

MacGregor, R.J. (1987) *Neural and Brain Modelling*, Academic Press, San Diego.

Mahowald, M. and Douglas, R. (1991) "A silicon neuron," *Nature*, Vol. 354, pp. 515-518.

Mead, C.A. (1989) *Analog VLSI and Neural Systems*, Addison-Wesley, Reading MA.

Meddis, R. and Hewitt, M. (1993) "Computational modeling of cochlear nucleus functioning," in Merchan, et al. (eds), *The Mammalian Cochlear Nucleï: Organisation and Function*, Plenum Press, New York.

Merchan, M.A., Juiz, J.M., Godfrey, D.A., and Mugaini, E. (eds) (1993), *The Mammalian Cochlear Nucleï: Organisation and Function*, Plenum Press, New York.

Mortara, A. (1995) *Communication Techniques for Analog VLSI Perceptive Systems*, Ph.D. Thesis, Ecole Polytechnique Fédérale, Lausanne.

Oertel, D. (1983) "Synaptic responses and electrical properties of cells in brain slices of the mouse anteroventral cochlear nucleus," *Journal of Neuroscience*, Vol. 3, pp. 2043-2053.

Popper, A.N. and Fay, R.R. (eds) (1992), *The Mammalian Auditory Pathway: Neurophysiology*, Springer-Verlag, New York.

Rasche, C., Douglas, R.J., and Mahowald, M. (1997) "Characterization of a Pyramidal Silicon Neuron," *Proceedings of the EWNS'97*.

Rhode, W.S. and Greenberg, S. (1992) "Physiology of the Cochlear Nuclei" in Popper and Fay, (eds), *The Mammalian Auditory Pathway: Neurophysiology*, Springer-Verlag, New York, pp. 94-152.

Rhode, W.S. and Greenberg, S. (1994) "Encoding of Amplitude Modulation in the Cochlear Nucleus of the Cat," *Journal of Neurophysiology*, Vol. 71, No. 5, pp. 1797-1825.

Sarpeshkar, R., Watts, L., and Mead, C. (1992) "Refractory Neuron Circuits," *CNS Technical Report*, No. CNS-TR-92-08, California Institute of Technology, Pasadena, CA.

Tuckwell, H.C. (1988) *Introduction to theoretical neurobiology*, Cambridge University Press, Cambridge UK.

Vittoz, E.A. (1994) "Analog VLSI signal processing : why, where, and how," *Journal of VLSI Signal Processing*, Vol. 8, pp. 27-44.

Webster, D.B., Popper, A.N., and Fay, R.R. (eds) (1992), *The Mammalian Auditory Pathway: Neuroanatomy*, Springer-Verlag, New York.

6. Periodicity Extraction

6.1 Introduction

While the simulation of the activity of individual neural units or the signal processing in the cochlea is of scientific interest in its own right, the goal of my research was to build and understand the signal processing of a system containing cochlear pre-processing followed by further processing by large ensembles of spiking neurons. In fact, VLSI architectures are probably not the best way of modelling activity in single neurons. Too many compromises are required to make the circuit design problem tractable. The flexibility of computer programming is much better suited to this purpose. However the massive parallel processing capability of VLSI circuits and their real-time functioning will make them the preferred medium for studying large numbers of interacting neurons.

Anatomy can supply us with some information concerning the connectivity between nuclei in the auditory brainstem. To a more limited extent it can even identify which specific cell types within a nucleus project to which other specific cell types within another nucleus, although this information is sometimes speculative. Anatomical techniques also exist for identifying whether certain connections are excitatory or inhibitory. The job of the modeller is to propose and evaluate anatomically plausible circuits which could give rise to the observed physiological response of the ensemble of cells to acoustic stimulation. Only then will we be able to understand the fine points of the signal processing performed by this ensemble of cells. This necessarily involves models that begin with acoustic input and follow the processing through all of the peripheral stages as well as the neural responses, and this is the reason why the electronic building blocks have been created. Thus, although the anatomical and physiological picture is far from complete, we have enough information to begin the process of generating some simple models to mimic some known circuits in the auditory brainstem. The electronic building blocks are far from being perfect, but they already allow us to start building electronic models of some simple circuits in the auditory brainstem. Only by building operational neural systems with the building blocks we can prove their usefulness.

95

In this chapter and the next I will present two neural systems that use the building blocks to extract periodicity. The first system extracts the periodicity that corresponds to the frequency of a pure tone or to the carrier of an amplitude modulated tone. The second system extracts the periodicity of the envelope of the signal. In both systems, the periodicity of the signal or envelope corresponds to the repetition rate of the peaks of the waveform amplitude.

6.2 Biological background

The inner hair cell measurements in chapter 2 show that for frequencies above about 4 kHz the inner hair cell receptor potential has lost its a.c. component almost completely; no information can thus be available in the brain about the temporal properties of the input waveform. For these frequencies, the spectrum of the sound can only be coded by the position of the active auditory nerve fibres and the mean spiking rate of these fibres. On the other hand, at lower frequencies the inner hair cell receptor potential has a noticeable a.c. component and the auditory nerve fibres spike with a probability roughly proportional to the half-wave rectified basilar membrane velocity waveform at the cell's position (Palmer and Russell, 1986). Therefore, for these lower frequencies, the spectrum of the input waveform, is not only encoded in the form of an average spiking rate of different fibres along the cochlea (place coding), but also in the periodicity of spiking of the individual auditory nerve fibres.

It has been shown that this periodicity information is a much more robust cue than the spatial distribution of average firing rates. Some periodicity information can already be detected at intensities 20 dB below the intensity needed to obtain a change in average rate (Evans, 1982). Periodicity information is retained at intensities in the range of 60-90 dB SPL, for which the average rate of the majority of the auditory nerve fibres is saturated (Rose, et al., 1971). Moreover, the positions of the fibres responding best to a given frequency move with changing sound intensity, whereas the periodicity information remains constant. Furthermore, the frequency selectivity of a given fibre's spiking rate is drastically reduced at medium and high sound intensities. The robustness of periodicity information makes it likely that the brain actually uses this information, although no direct evidence for this has been found.

6.3 Modelling periodicity extraction

6.3.1 Existing models

We shall first look at some existing models for extracting periodicity in this section. In the next section I will then describe the improvement of one of these models and its implementation.

Several models have been proposed that extract periodicity information using the phase encoding of fibres connected to the same inner hair cell (Delgutte, 1984; Lyon, 1984, Sachs and Young, 1988; Seneff, 1988). Others propose using synchronicity of firing on auditory nerve fibres connected to different inner hair cells (Deng, et al, 1988; Ghitza, 1988; Shamma, 1988). However, to date no evidence has been found in support of any of these models and they remain therefore hypothetical.

The simplest of the phase encoding schemes compares the output of the cochlea at a given position with a delayed version of itself. It is easy to see that for pure tones, the comparison:

$$\sin(2 \pi f t) = \sin(2 \pi f (t - \Delta)) \tag{53}$$

is only true for:

$$f = N / \Delta \tag{54}$$

where N is an integer. Thus only for frequencies that are a multiple of $1/\Delta$, the signals are in perfect synchrony and thus perfectly correlated. We can adapt the delay Δ to each cochlear output, so that $1/\Delta$ equals the best frequency of that cochlear output. In this case higher multiples of $1/\Delta$ will be suppressed due to the very steep cut-off of the cochlear filters for frequencies above the best frequency. Each synchronicity detector will then only be sensitive to the best frequency of the filter to which it is connected. If we code the signals by a spike train, with one spike per period at a fixed phase, it becomes a very simple operation to detect this synchronicity. A simple digital AND operator will be enough to detect overlap between two spikes. These spikes will overlap perfectly when $f = 1/\Delta$, but some overlap will still be present for frequencies close to $1/\Delta$, since the spikes have a finite width. The bandwidth of the AND output can thus be controlled by the spike width.

It is possible to create a silicon implementation of this scheme using the artificial cochlea and IHC circuit of Chapter 3, the spiking neuron of Chapter 5 and the axon hillock circuit to create the delays. A similar chip has been developed by John Lazzaro (1991) and functioned correctly.

A disadvantage of this scheme is the fact that the delay associated with a cochlear output has to be matched to the inverse of the best frequency of that cochlear output. For a cochlea whose best frequency changes exponentially with filter number in the cascade from 4 kHz (the upper range of phase locking on the auditory nerve) to 100 Hz, we will have to create delays that range from 0.25 ms to 10 ms. In the brain, such a large variation in delays is unlikely to be provided by an axonal delay circuit because that would require an excessive large variation in axon length.

6.3.2 *Improved model*

A possible solution comes from the observation that the phase of a pure tone of a given frequency on the basilar membrane increases from base to apex, and the phase changes rapidly around the best frequency. For the silicon cochlea, the filter cascade also functions as a delay line, and each filter adds a delay which corresponds to $\pi/2$ at the cut-off frequency of that filter. If we monitor the output of filter i and filter i-4, under the assumption that the filters have exactly the same cut-off frequency, then we have a delay corresponding to 2π at the cut-off frequency.
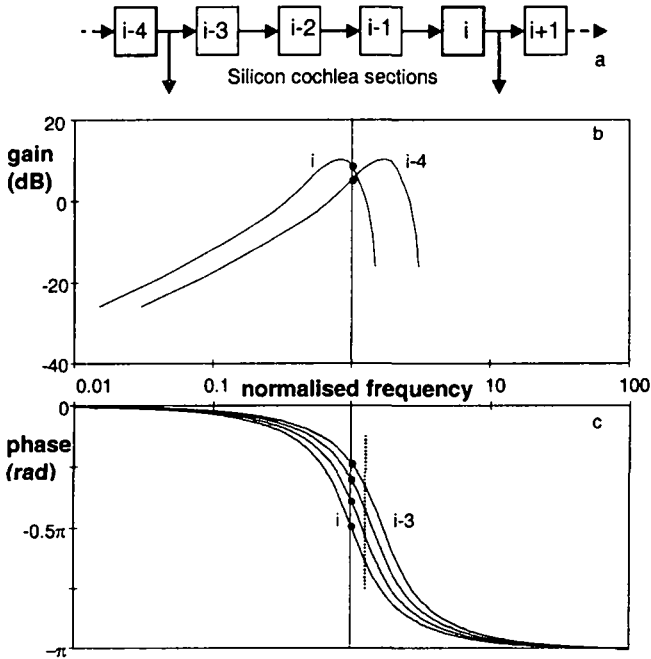
Figure 6.1 a) Accumulated gain at output i and i-4; b) phase curves of the individual stages between output i and output i-4.

In reality, the filters along the cochlea will have different cut-off frequencies, as shown in Figure 6.1. This figure shows the *accumulated* gain at the outputs i and i-4, and the delay added by each *individual* filter between these two outputs as a function of frequency normalised to the cut-off frequency of filter i. The vertical line represents this cut-off frequency, and we can see that only filter i adds a delay of $\pi/2$, and the other filters add less. However, if we move the vertical line to the right, the delay added by each filter will increase relatively

quickly, and at some frequency slightly higher than the cut-off frequency of filter i, the sum of the delays will become 2π (dashed line). At this frequency neither filter i nor i-4 has maximum gain, but if the cut-off frequency of both filters is not too different, e.g., four filters per octave or more, the gain will still be high enough for both filters at the correlator frequency to yield output signals with reasonable amplitudes. For the same reason, we should use the minimum number of filters between the two outputs that we correlate.

It is possible to obtain a phase difference of 2π with three filters already, but only at a frequency well above the cut-off frequency of each of the filters. In this case the response has a very small amplitude. For this reason I have chosen to use four filters between the two outputs. The phase difference between two such outputs is given by:

$$\Delta\phi = \sum_{x=i-4}^{x=i} \arctan\left(\frac{-f/fc_x}{Q\,(1-(f/fc_x)^2)}\right) \tag{55}$$

where fc_x is the cut-off frequency of filter x. Actually the arctan function jumps from $-\pi/2$ to $\pi/2$ at f=fc, and we will have to subtract π from the output of the function whenever f>fc, to keep the function continuous. The equation shows that $\Delta\phi$ is not only a function of frequency, but also of the quality factor of the filters. Again, when the four filters have very similar cut-off frequency, the signal frequency that will yield a $\Delta\phi = -2\pi$ is very close to the cut-off frequency of these filters and the term $(1-(f/fc)^2)$ will become very small, so that $\Delta\phi$ becomes almost independent of Q.
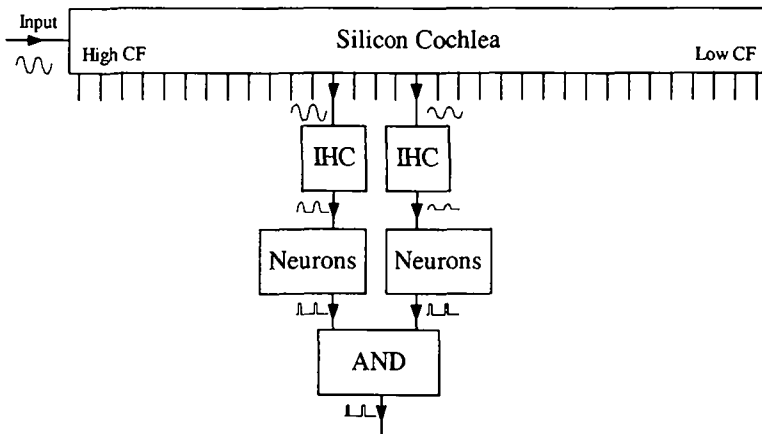
Figure 6.2 Implementation of the periodicity extraction model.

6.4 Implementation with the building blocks

I have implemented this system using my building blocks as shown in Figure 6.2. The silicon cochlea is used to filter and delay the signal, and has been adjusted so that the cut-off frequency decreases by one octave every twenty stages, so that the cut-off frequencies of neighbouring filters are almost equal. The IHC circuit is used to half-wave rectify the signal, but its low-pass filter function has been disabled in the first series of experiments, in order to avoid reduction of phase-locking with increasing signal frequency. The 32 spiking neurons on the neuron chip are used as 32 spiral ganglion cells that take their input from the same inner hair cell.

To simplify the system in the first tests, I have adapted the settings of the neurons on a chip so that they behave as a single neuron that always gives exactly one spike per period of the input signal. A digital AND gate is used to compare the output spikes of the two chips, and the spike rate at the output of the AND gate is the measure of activity used. The spike width of the neurons will control the frequency selectivity of the system. The digital AND gate only changes the output spike width when spikes overlap partially, but the spike rate will stay the same; measured this way, the temporal resolution of the system is defined by the spike width. This is not a fully realistic model with respect to biology.

Auditory nerve spikes arriving on a real nerve cell performing the synchrony detection will yield excitatory post synaptic potentials (EPSPs) which have a clear maximum at a given time after the arrival of the spike, as shown in Figure 5.13a. Because of this temporal maximum, two EPSPs integrated on the cell's membrane capacitance will yield a maximum voltage change when both spikes arrive together, and a smaller voltage change when spikes overlap only partially. The biological cell could use this fact to create a temporal resolution better than the width of a spike. In the following experiments the neurons have therefore been set to a 100 μs spike width, about five times shorter than the width of auditory nerve spikes, to reflect a hypothetical enhanced temporal resolution in the biological system.

6.5 Test results

6.5.1 Output at different filters

In the first experiment, I have measured the number of spikes per second at the output of the AND gate as a function of input frequency and using different cochlear filters. The silicon cochlea has been adjusted to have twenty filter sections per octave; I have measured twelve filter combinations, each combining a filter output with the output of a filter four sections earlier in the

cascade. The best frequency of the filter with the lowest best frequency of the two filters ranged from 200 Hz to 880 Hz. I have measured the output every four sections. The results are shown in Figure 6.3a
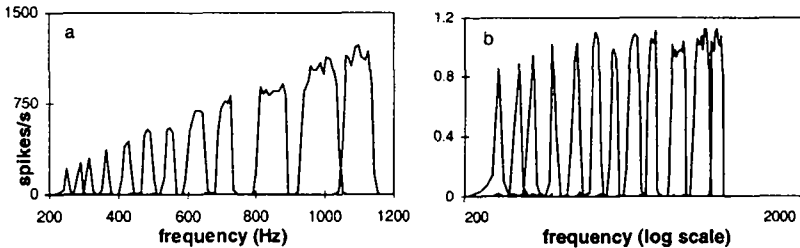

Figure 6.3 a) measured output rate at different cochlear positions, and b) spike rate normalised to best input frequency, plotted on a log frequency scale.

The maximum spike rate increases approximately linearly with frequency; this is to be expected, since we will have approximately one spike per signal period. Furthermore the best response frequencies of the filters sensitive to higher frequencies are further apart, due to the exponential scaling of the frequencies along the cochlea. Finally, a given time delay corresponds to a larger phase delay for the higher frequencies, so that the absolute bandwidth of the coincidence detectors, i.e., the range of input frequencies to which they respond, is larger. When we normalise the spike rate, and plot the curves on a logarithmic frequency scale, as in Figure 6.3b, we see that the best frequencies of the correlators follow the exponential scaling of the best frequencies of the cochlear filters, and that the relative bandwidth is fairly constant.

It is also interesting to note that although I use only one out of every four silicon cochlear outputs, I have access to one out of every two cochlear outputs on the cochlea chip. So in between each curve in Figure 6.3a I can measure one other curve, and this would just about cover all the possible input frequencies, with little overlap between the selectivity curves of the correlators. A shorter spike width would yield more selective correlator outputs, but input frequencies that fall between two peaks would not elicit a response at all. For a setting with 20 stages per octave of the silicon cochlea, Figure 6.3 thus represents the situation with the most frequency selective settings for the correlator that covers all input frequencies.

6.5.2 Amplitude dependence

Using the same settings as in the previous experiment, I have measured the output spike rate of the system for different input amplitudes, when connected to the cochlear filters with a 710 Hz and 810 Hz best frequency respectively. In principle, the amplitude of the input signal should have no effect on the output

of the system, since the system only uses phase information. However, this is only true if the spikes are always created at the same phase of the output signal of the cochlear filters, for instance at the peak, or the zero crossing. Figure 6.4 shows however that the resulting filter selectivity shifts to lower frequencies for higher intensity input signals.
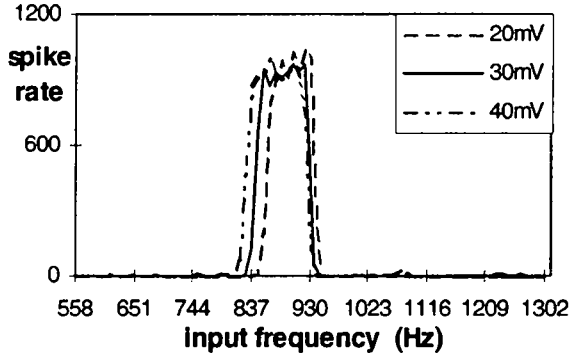

Figure 6.4 Frequency selectivity for different input intensities.

This is a result of the way the spikes are created on the neuron chip. The neurons have been adjusted to spike once per period, but the phase at which they spike with respect to the half-wave-rectified waveform depends only on the integration time T_I of the neuron, which is the time needed with a given input current to reach the threshold voltage V_T from the zero resting voltage. This time depends on the amplitude of the input current, which in turn is proportional to the amplitude of the input signal. Since the amplitude gain of the two cochlear filters used is not the same, the amplitude of the current input to the two neuron chips is different. Therefore, they do not spike at the same phase with respect to their respective input waveforms. The output spike trains of the two neuron chips are therefore perfectly synchronous, not when $\Delta\phi = -2\pi$ (as defined in and below equation (55)) but when $\Delta\phi + \Delta\phi_s = -2\pi$, where $\Delta\phi_s$ is the difference in spiking phase with respect to the local IHC output. Even though $\Delta\phi$ is independent of the input amplitude, $\Delta\phi_s$ isn't and this changes the frequency for which the two spike trains are in synchrony.

To analyse the dependence of $\Delta\phi_s$ to the input amplitude, we will start to determine the phase at which a given neuron chip spikes with respect to its input signal. We can write for the integration of the input current on the membrane capacitance:

$$I = G\,I_0\,\sin(2\,\pi\,f\,t) = C\,\frac{dV}{dt} \tag{56}$$

where G is the accumulated amplitude gain between the input signal of the cochlea and the output of the filter used. Furthermore, I_0 is the amplitude of the IHC output current when G is one, and C is the membrane capacitance. When we integrate both sides of the equation for the duration of the integration time T_I, we obtain:

$$\int_0^{T_I} G\, I_0 \sin(2\,\pi\, f\, t)\, dt = C\, V_\theta \qquad (57)$$

or:

$$\frac{G\, I_0}{2\,\pi\, f}\, (1 - \cos(2\,\pi\, f\, T_I)) = C\, V_\theta \qquad (58)$$

which we can rewrite as:

$$2\,\pi\, f\, T_I = \arccos\left(1 - \frac{2\,\pi\, f\, C\, V_T}{G\, I_0}\right) = \phi_S \qquad (59)$$

For values of its argument between 0 and 1, the arccos function decreases monotonously. The interesting section is from $\pi/2$ to 0, which are the phases of the positive half cycle of the input current. This is the only phase range allowed for a one-spike-per-period setting. This means that (59) can only be used when $GI_0 \geq 2\pi f C V_T$ and that ϕ_S becomes almost zero for very large input signals. For the difference in spiking phase of the two neuron chips with inputs (via the IHC circuit) from different cochlear filter outputs we can thus write:

$$\Delta\phi_S = \phi_{S1} - \phi_{S2} = \arccos\left(1 - \frac{2\,\pi\, f\, C\, V_T}{G_1\, I_0}\right) - \arccos\left(1 - \frac{2\,\pi\, f\, C\, V_T}{G_2\, I_0}\right) \quad (60)$$

where G_1 and G_2 are the amplitude gains of the two respective cochlear filters. With this equation we can calculate that $\Delta\phi_S$ is positive and decreases for increasing input amplitude when G_2 is larger than G_1. This is the case in the current example, because the best frequency (BF) of the first filter is 710 Hz and 810 Hz for the second filter; the frequencies which yield synchronisation are above 810 Hz, so that the 810 Hz BF filter will have higher gain than the 710 Hz BF filter. If $\Delta\phi_S$ decreases with intensity, then $\Delta\phi$ will have to increase, i.e., become less negative, in order to keep their sum constant at -2π and we can see in Figure 6.1 that $\Delta\phi$ decreases with decreasing signal frequency.

The analysis above explains why the frequency selectivity of the system shifts to lower frequencies with increasing intensity, but this is an artefact of the spike generation used to simplify the system. On the auditory nerve, spikes

arrive with a probability roughly proportional to the half-wave rectified waveform. The most probable phase for a spike is therefore always at the maximum of the waveform, independent of intensity. In such a system, the frequency selectivity will therefore be independent of amplitude. A second advantage of coding (at least half of) the waveform in spike probability is that it does not assume that the input waveform is sinusoidal. Coding a waveform with one spike per period can only code the frequency and phase of a waveform, but not its form.

6.5.3 Amplitude modulated signals

To test the model with a more complex waveform, I have used a 930 Hz sine wave 100% amplitude-modulated at 200 Hz generated on a computer. Because of the way the input frequency is varied on the computer during playback — by playing the whole waveform a certain percentage slower or faster — the actual modulation frequency changes with the same factor as the carrier frequency. The results of this test are shown in Figure 6.5 for three different input amplitudes.
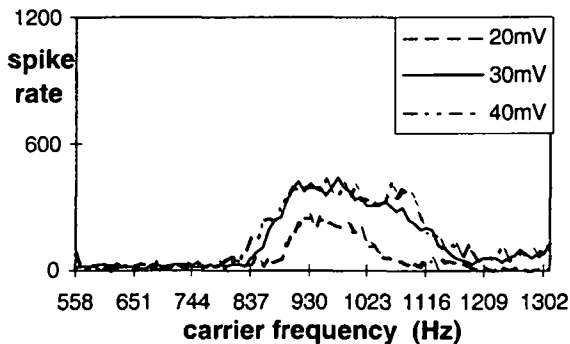
Figure 6.5 Frequency selectivity for different input intensities with AM input.

Compared to the measurements in Figure 6.4, we see that the filter is less selective and centred at a higher input frequency. The shift towards a higher frequency can be explained by the fact that the average input amplitude of an amplitude modulated signal is lower than in a pure tone with the same maximum amplitude. Furthermore, the amplitude of the positive half-cycle of the output of the IHC circuit changes from cycle to cycle because of the amplitude modulation. We have seen that the amplitude of the input signal changes the frequency for which the two spike trains are synchronous, which means that the frequency which yields the best response changes from cycle to cycle with a periodicity equal to the modulation frequency. This introduces a sort of "roaming" of the frequencies in the input signal, effectively reducing the

selectivity of the filters. Finally, because of the 100% depth of the amplitude modulation, the amplitude of the input will be too low during some cycles to create a spike, which therefore reduces the total number of spikes which can coincide.

6.5.4 Periodicity vs. spectral content

Figure 6.5 shows that this model detects periodicity and not spectral content. The spectrum of a 930 Hz pure tone 100% amplitude modulated at 200 Hz contains, apart from a 930 Hz carrier component, components both at 730 Hz and 1130 Hz, with half the amplitude of the carrier component. When the speed of the waveform playback is varied so that the carrier frequency is either 765 Hz or 1185 Hz, one of these spectral side bands will be at 930 Hz, but the system does not respond to these spectral components. This is explained by the fact that the periodicity of the zero crossings, and thus of the positive half cycles of the IHC output, is always equal to the carrier frequency.
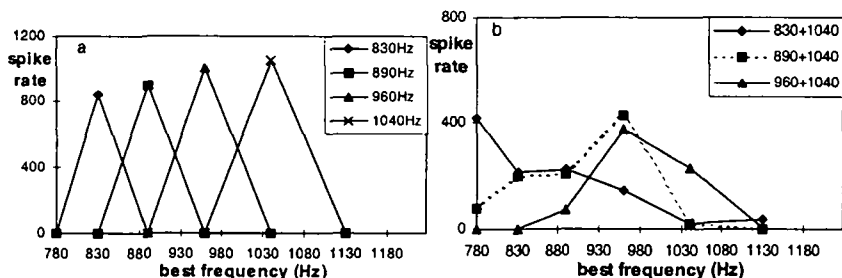
Figure 6.6 Response of six correlators with different best frequencies to a) four different pure tones, and b) three different mixes of pure tones.

The difference between periodicity detection and spectral content extraction is even clearer in Figure 6.6. This figure is different from the previous ones in the fact that it is not the input frequency that varies on the x-axis, but the best frequency of the synchronicity detector; the different curves are for different input signals. Figure 6.6a shows the response for four different pure tones at the best frequencies of the four middle synchronicity detectors. This shows that effectively only one detector reacts for each input signal, and thus signals the presence of an input signal with a periodicity that is within its pass-band.

However, when we mix two of such pure tones at equal amplitude, we obtain the result of Figure 6.6b, which shows that the two spectral components are not separated, and sometimes not even detected. Again this can be explained by the periodicity of the input signal. When the intervals between for instance two successive peaks or zero-crossings are measured, we see that these intervals vary from cycle to cycle. Therefore, within the range of delays we are interested

in, it is impossible to find a single delay different from zero that will line up all the peaks in the two cochlear outputs. There are several delays that will line up some of the peaks, making certain synchronicity detectors respond weakly; these delays do not necessarily correspond to the inverse of the pure tone frequencies used in the mix. For instance in the mix of a 1040 Hz and a 830 Hz sine wave, we find time intervals between successive peaks that correspond to instantaneous frequencies in the range of 750-870 Hz, and not a single one in the 1000 Hz region. In Figure 6.6 we can see that indeed only the detectors in the low frequency range respond.

6.5.5 Noise

The fact that the system is quite a good periodicity detector is shown by the measurements of Figure 6.7. In these measurements, white noise has been added with different gains to a 930 Hz pure tone with a 20 mV amplitude to create different signal to noise ratios (SNRs). The noise has been generated with the same maximum amplitude as the pure tone and with a 11025 Hz sample frequency. The addition of a small amount of noise to the signal results in phase jitter in the spike trains, which increases the range of frequencies which are able to elicit coinciding spikes, but also changes the flat topped frequency selectivity curve to a pointed one, and thus renders the best frequency more visible. When the noise increases, the average number of spikes in the pass-band decreases, and the average number of spikes outside the pass-band increases. With a noise gain as high as 2, there still is a clear difference between the average spike rate in the pass-band and outside. Finally, at a noise gain of four, the average spike rate is constant, except that there is a noticeable spike rate increase at best frequency only. In this case the signal does not respond strongly to a signal at best frequency, but it does respond very selectively!
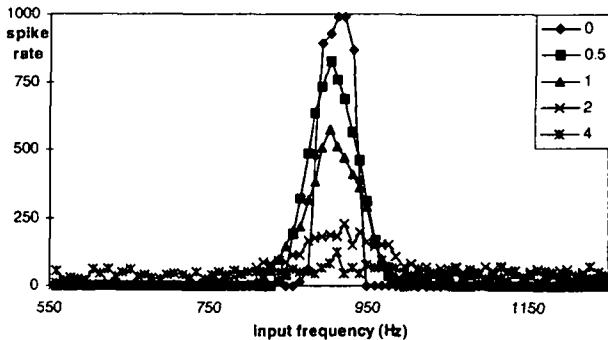
Figure 6.7 Frequency selectivity for a 930Hz pure tone with white noise added with different gains. The noise gain is given in the legend.

6.5.6 Temporal behaviour

Traditional band-pass filters with a very high quality factor (Q) can also yield a narrow pass-band, but their step response takes about 1.5Q cycles at the centre frequency to settle. The periodicity selectivity of the synchronicity detector shown in Figure 6.7 corresponds to a quality factor of 14 at 30dB SNR; a traditional band-pass filter would take about 21 cycles of the 930Hz input signal to settle. Figure 6.8 shows the temporal aspect of the synchronicity detection in our system. The top trace in this figure shows the output of the cochlear filter with the highest best frequency (index i-4 in Figure 6.1) and the spikes generated based on this output. The second trace shows the same for the output of the cochlear filter with the lower best frequency (index i in Figure 6.1). The third trace shows the output of the AND gate with the above inputs, which are slightly above its best periodicity. Coincidences are detected at the onset of the tone, even when it is not of the correct periodicity, but only for the first one or two cycles. The bottom trace shows the output of the AND gate for an input at best frequency. The system thus responds to the presence of a pure tone of the correct periodicity after only a few cycles, independent of the filters selectivity.
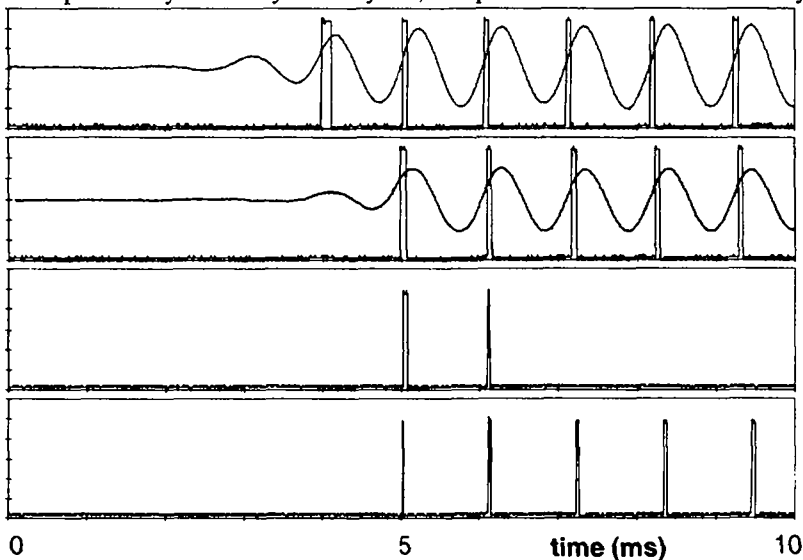
Figure 6.8 Oscilloscope traces of the temporal aspect of synchronicity detection. See text for details. The vertical scale is 20mV per square for the cochlear outputs, the spikes are 5V in amplitude.

To show this more dramatically, I have reduced the spike width to 10 µs, to obtain a high periodicity selectivity as shown in Figure 6.9. The bandwidth of this filter is only 20 Hz at 930 Hz, equivalent to a quality factor of 46.5. A

traditional filter with such a quality factor would only settle 70 cycles after the onset of the signal, whereas the periodicity detector still settles after the first two cycles, as shown in Figure 6.10. Note that, because of the very short spikes, coincidences are not detected at every cycle, reflected by a lower maximum spike rate in Figure 6.9.
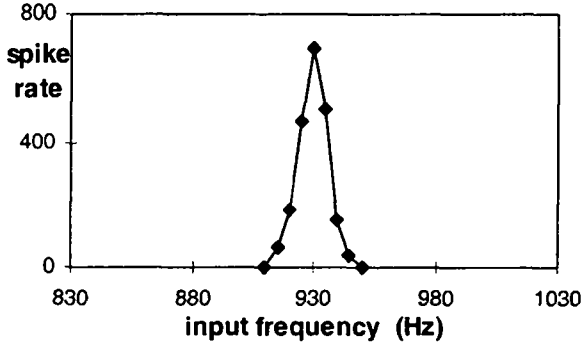

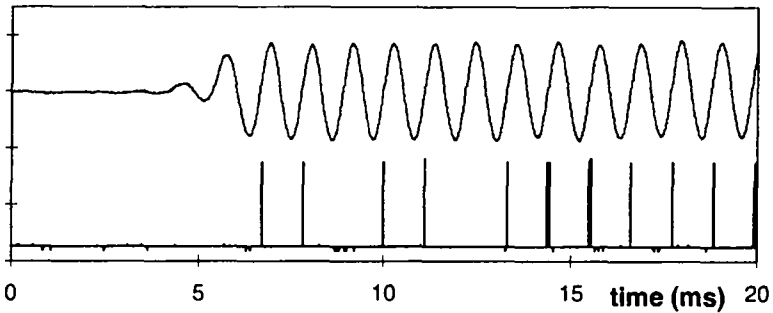**Figure 6.9** Frequency selectivity with a 10μs spike width.

Figure 6.10 Cochlear output (top, 40 mV scale) and coincidences (bottom) for a signal at best frequency with a 10 μs spike width.

We can compare this result with the response of a classic RLC band-pass filter with a 930 Hz centre frequency and a quality factor of 46.5 as shown in Figure 6.11 and Figure 6.12. After 18 cycles of the input signal, the output of the band-pass filter has only reached 69% of its final value.
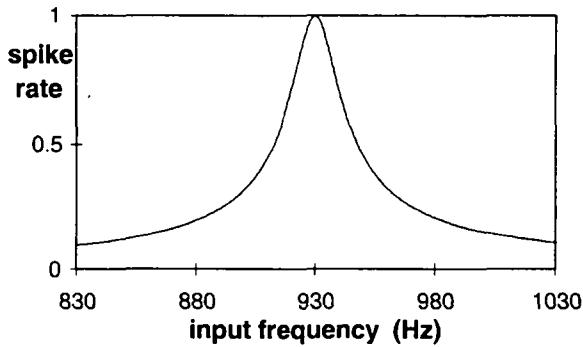
Figure 6.11 Simulated frequency selectivity of the RLC band-pass filter.
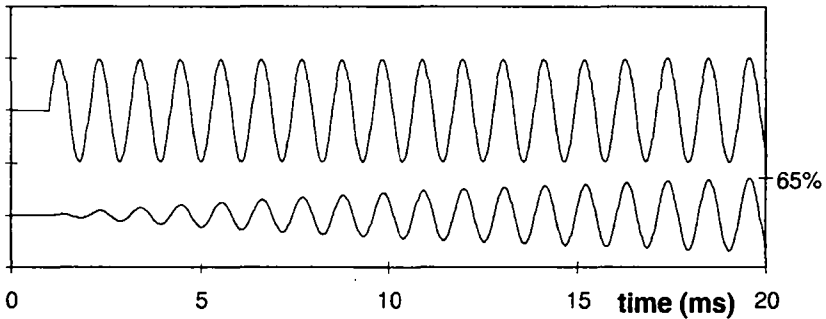
Figure 6.12 Simulated transient response of the RLC band-pass filter. Scale units are 40 mV

6.5.7 Biological spike distribution

In the previous sections we simplified the model to use one spike per period in order to understand the principle behind the periodicity detection. However, we have seen that the implementation using the 32 neurons behaving as a single neuron (to give one spike per period) leads to a shift in best periodicity with changing amplitude, because the phase at which the 'single neuron' spikes changes with intensity. Next, we change the settings to be more realistic, so that none of the 32 neurons can spike at each period, and we reduce the output gain of the IHC circuit so that the neurons receive less signal current, and thus have a lower input SNR. The resolving spike distribution resembles the spike distribution on the auditory nerve more. This is shown in Figure 6.13 for a group of 32 neurons stimulated by and IHC circuit connected to a single cochlear output. The bottom trace shows the sum of spikes over the 32 neurons.
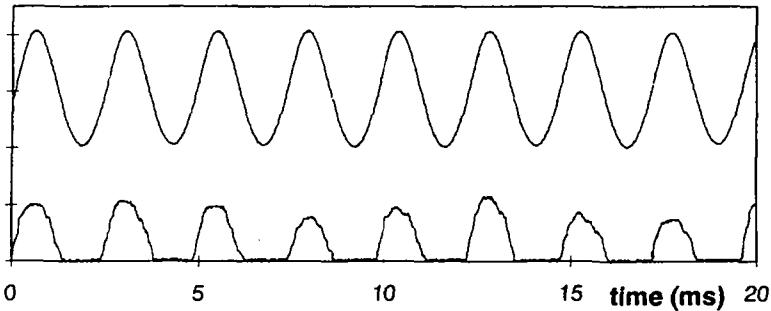
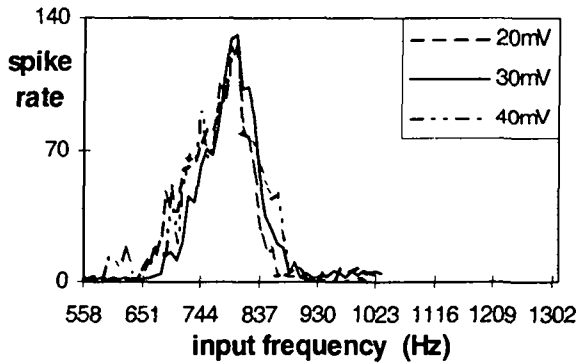**Figure 6.13** Cochlear output (top) and population average of the AN spikes (bottom).

Figure 6.14 Periodicity selectivity with auditory nerve like spike distribution.

When we use spike distributions as shown in Figure 6.13 and repeat the pure-tone detection experiment of Figure 6.4 at different input intensities, we obtain the periodicity selectivity curve of Figure 6.14. Indeed, in this case, the best periodicity does not change; the curves are remarkably independent of input intensity. The selectivity curve is about twice as wide at the base as the ones in Figure 6.4. Furthermore, we do not get coincidences at each period, so the maximum number of spikes is lower. Finally the slopes of the selectivity curve rise and fall much more gradually, just as in the added-noise case in Figure 6.7. This also means that we can easily increase the selectivity of these curves by setting a higher threshold, e.g., discarding spike rates below 70 spikes per second. Because of the steep slopes in Figure 6.4 such an operation would hardly increase the selectivity for that case.

This first application of the building blocks clearly demonstrates the advantages of using spikes in the detection of periodicity and that simplification of the system might help to better understand the principle behind the operation;

it is also clear that simplification can also introduce artefacts which are not present in a system which mimics the biology more closely.

6.6 Summary

In this chapter I have presented a neural system implemented with the building blocks from the previous chapters. The system uses the delay between the outputs at two points along the cochlea to detect a periodicity in the input signal.

An especially useful property of the cochlea is that the delay between two points with a fixed distance between them corresponds to a full period at a frequency that scales in the same way as the best frequency scales along the cochlea, i.e., it decreases exponentially.

If we always create spikes at the same phase of the output signal at each filter, or simply have the highest spiking probability for the maximum instantaneous amplitude of the output signal, then both outputs will only have synchronous spikes for a certain periodicity, and we can easily detect this synchronicity with coincidence detectors.

This system offers a way to obtain very selective filters using spikes. Even though they react to a very narrow range of periodicities, these filters are able to react after only a few periods. Furthermore, the range of periodicities it responds to can be made independent of input intensity, which is not the case with the cochlear output itself.

6.7 References for chapter 6

Delgutte, B. (1984) "Speech coding in the auditory nerve: II. Processing schemses for vowel like sounds," *Journal of the Acoustical Society of America*, Vol. 75, no. 3, pp. 879-886.

Deng, L., Geisler, C.D., and Greenberg, S. (1988) "A composite model of the auditory periphery for the processing of speech," *Journal of Phonetics*, Vol. 16, pp. 93-108.

Evans, E.F. (1982), "Functional anatomy of the auditory system," in Barlow and Mollon (editors), *The Senses*, Cambridge University Press, Cambridge, pp. 251-306.

Ghitza, O. (1988) "Temporal non-place information in the auditory-nerve firing patterns as a front end for speech recognition in a noisy environment," *Journal of Phonetics*, Vol. 16, pp. 109-123.

Lazzaro, J. (1991) "A silicon model of an auditory neural representation of spectral shape." *IEEE Journal of Solid-State Circuits*, Vol. 26, No. 5, pp. 772-777.

Lyon, R.F. (1984) "Computational models of neural auditory processing," in *Proceedings 1984 IEEE ICASSP*, San Diego.

Palmer, A.R. and Russell, I.J. (1986) "Phase locking in the cochlear nerve of the guinea pig and its relation to the receptor potential of inner hair cells," *Hearing Research*, Vol. 24, pp. 1-15.

Rose, J.E., Hind, J.E., Anderson, D.J., and Brugge, J.F. (1971) "Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey," *Journal of Neurophysiology*, Vol. 34, pp. 685-699.

Sachs, M.B. and Young, E.D. (1980) "Effects of nonlinearities on speech encoding in the auditory nerve," *Journal of the Acoustical Society of America*, Vol. 68, no. 3, pp. 858-875.

Shamma, S. (1988) "The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives," *Journal of Phonetics*, Vol. 16, pp. 77-91.

Seneff, S. (1988) "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, Vol. 16, pp. 55-76.

.

7. Envelope periodicity extraction

7.1 Biological background

Another form of periodicity detection that I have tried to model with the electronic building blocks is envelope periodicity detection, i.e., the extraction of the repetition rate of the peaks in the waveform envelope. This can be used to extract the fundamental of a harmonic complex, or the modulation frequency of an amplitude-modulated signal.

Amplitude Modulated (AM) sounds are of particular interest because this class of signals includes music and voiced speech, where the pitch of the sound is related to the rate of modulation. It has recently been discovered that stellate cells can be entrained to AM sounds; that is, they fire in synchrony with the peaks of the modulation envelope (Frisina, et al., 1990; Kim, et al., 1990). Sustained chopper cells, in particular, are entrained by a limited range of AM frequencies only. In other words, they have a band-pass transfer function for amplitude modulation which suggests they might be involved in pitch perception.

It is important to note that in most situations the firing rate of the chopper cell is only minimally affected by changes in the frequency of modulation of the signal. It is mainly the coherence of the cell's activity with respect to the acoustic stimulus that is influenced. Because different cells respond to different AM frequencies, we infer that pitch is represented spatially in the cochlear nucleus. Computer studies that simulate sustained chopping activity in response to AM signals suggest that the preferred rate of modulation is limited by the natural rate of chopping of the cell (Kim, et al., 1990). This, in turn, is related to intrinsic properties of the cell such as membrane capacitance or the time constant of recovery of potassium channels.

We can however go further than this, because it is known that some cells in the inferior colliculus (IC) also respond selectively to AM rates, but this time by increasing the *firing rate* of the cell as the rate of modulation approaches the cell's preferred frequency. It has been shown that this effect can be replicated by directing the output of a number of simulated cochlear nucleus (CN) chopper cells (with the same best modulation frequency) into a coincidence-detector

type of neuron (Hewitt and Meddis, 1994). This latter cell responds with a high firing rate only when the inputs from the CN are synchronised. In the case of sustained chopper cells, this condition is met only when they are all driven by an AM acoustic stimulus at their preferred AM frequency. This idea is supported (but not proven) by anatomical studies showing that some sustained chopper cells in the CN have excitatory projections to the IC (Webster, et al., 1992).
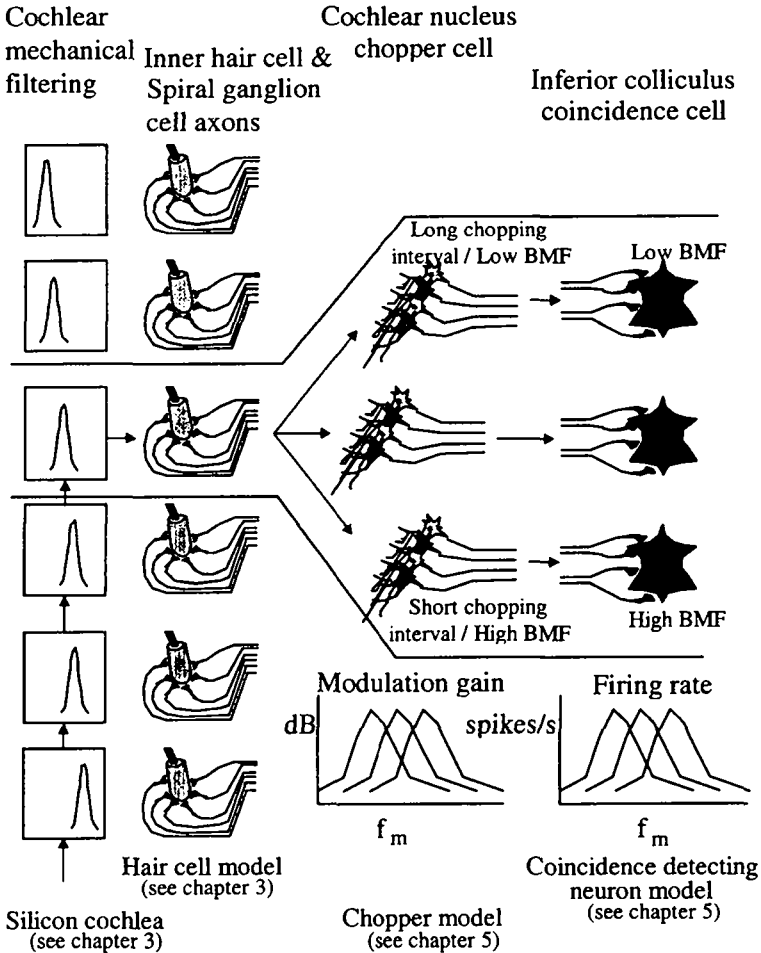
Cochlear
mechanical   Inner hair cell &   Cochlear nucleus
filtering   Spiral ganglion   chopper cell
              cell axons   Inferior colliculus
                                                  coincidence cell

Figure 7.1 Neural circuit for Amplitude Modulation sensitivity in the Inferior Colliculus proposed by Hewitt and Meddis (1994). BMF = best modulation frequency.

This theory of amplitude modulation sensitivity could be tested through extensive use of animal recordings but a good model of this process would allow us to explore the complex implications of the theory more conveniently and humanely. Hewitt and Meddis (1992,1994) have created computer models, which simulate the sensitivity of single choppers in the cochlear nucleus and of single coincidence-detecting neurons in the inferior colliculus. Although these models could in theory be extended to model multiple neurons functioning in parallel, the very long simulation times needed would make this very hard to use. Exploration of the model will be a lot easier once a hardware model has been constructed.

7.2 Amplitude modulation sensitivity model

7.2.1 Description of the model

Figure 7.1 shows the model as originally developed by Hewitt and Meddis (1994). In this model the incoming sound is filtered by the cochlea. The output at each point along the cochlea is transduced into neural signals by inner hair cells. The inner hair cells are contacted by spiral ganglion cells, which generate the auditory nerve spikes. The spike generation is a stochastic process, with a spike probability proportional to the inner hair cell output.

Each chopper is contacted by 60 auditory nerve fibres, all with the same best frequency. In the computer simulation of the model the choppers are stimulated by different auditory nerve fibres, but all fibres originate at the same inner hair cell. The choppers are divided into groups according to their chopping interval. A group of about 30 choppers with the same chopping interval then contacts the coincidence-detector neurons.

Choppers with the same chopping interval will synchronise for a signal with an amplitude modulation component whose frequency is close to the inverse of the chopping interval. This synchronicity will cause the coincidence-detectors, stimulated by these choppers, to fire.

7.2.2 Chopper synchronisation

To understand why the choppers will synchronise for a certain amplitude modulation frequency, one has to look at the signal envelope (Figure 7.2), which contains temporal information on a time scale that can influence the spiking neurons. Because of the low-pass filtering of the IHC circuit, the 5 kHz carrier itself will not contain any temporal information that influences the spiking neuron in an important way.

Consider the case when the modulation frequency is similar to the chopping frequency (Figure 7.2). If a chopper spikes during the rising flank of the envelope, it will come out of its refractory period just before the next rising

flank of the envelope. However, if the driven chopping frequency is a bit too low, the chopper will come out of its refractory period a bit later. This means that the chopper starts integrating its input higher on the rising flank of the envelope and therefore receives a higher average stimulation. This in turn increases its chopping frequency. The rising flank of the stimulus envelope thus provides a form of negative feedback on the chopping frequency, which makes spiking on a certain point on the rising flank of the envelope a stable situation. With the same reasoning one can show that spiking on the falling flank is unstable. A group of similar choppers thus stabilise at about the same point on the rising flank; their spikes will thus coincide when the modulation frequency allows them to.
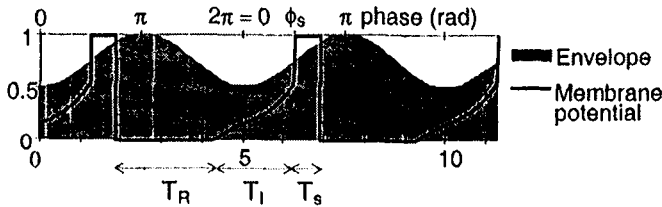
Figure 7.2 Spike generation for a chopper cell.

We can characterise this situation more precisely with some calculation. Consider a high-frequency input signal amplitude modulated at a frequency f_{mod} and modulation depth md, with maximum amplitude I_{max}, will create, after half wave rectification and low pass filtering by the IHC circuit, a stimulus current which can be approximated by the signal's envelope:

$$I_{stim} = I_{max}\left(1 - \frac{md}{2}\right) - I_{max}\left(\frac{md}{2}\right)Cos(2\pi f_{mod}t) \qquad (61)$$

This stimulus charges the membrane capacitance C_{mem} from zero volts to the threshold voltage V_{thres} of the neuron prior to spiking. If a neuron spikes at phase $\phi_s = 2\pi f_{mod}t_s$, we can calculate what the duration of the integration time T_I must have been, since:

$$I_{max}\left(1 - \frac{md}{2}\right) - I_{max}\left(\frac{md}{2}\right)Cos(2\pi f_{mod}t) = C_{mem}\frac{dV_{mem}}{dt} \qquad (62)$$

so that:

$$\int_{(\phi_s/2\pi f_{mod}) - T_I}^{\phi_s/2\pi f_{mod}}\left(I_{max}\left(1 - \frac{md}{2}\right) - I_{max}\left(\frac{md}{2}\right)Cos(2\pi f_{mod}t)\right)dt = C_{mem}V_{thres} \qquad (63)$$

which yields:

$$I_{max}\left(\left(1 - \frac{md}{2}\right)T_I - \frac{md(Sin(\phi_s) - Sin(\phi_s - 2\pi f_{mod}T_I))}{4\pi f_{mod}}\right) - C_{mem}V_{thres} = 0 \quad (64)$$

For a neuron to be phase locked at phase ϕ_s, it must always spike at ϕ_s, which means that $T_I + T_R + T_S = 1/f_{mod}$, where T_R is the refractory period of the neuron and T_S the spike width. I have plotted equation (61) and the numerical solution to (64) in Figure 7.3a and b, for md=0.5, I_{max}=15 nA, V_T=2 V, C_{mem}=15 pF, and f_{mod}=100 Hz.
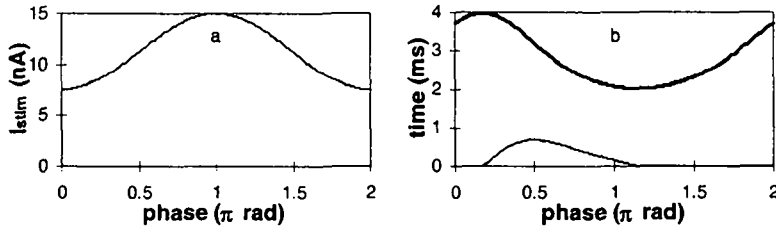
Figure 7.3 a) Stimulus current, and b) T_I (bold) and ΔT_R. See text for details.

When we know the spike width T_S, we can use equation (64) to calculate T_I for a given phase and thus what T_R should be to obtain stable phase locking at a given phase ϕ_s. As reasoned before, phase-locking of the neuron is only stable when $dT_I/d\phi_s$ is negative, so the T_R we calculate is only valid in this phase region. A different refractory period T_R will yield a different spiking phase ϕ_s, and we can calculate what the maximum allowable difference in T_R is between two otherwise identical neurons when we want their spikes to overlap. This ΔT_R curve is shown in Figure 7.3b for a spike width of 0.5 ms. We can see from this curve that in the above situation ΔT_R is largest for a phase of about 0.5π, and it will be at this phase that the largest group of neurons will be phase locked.
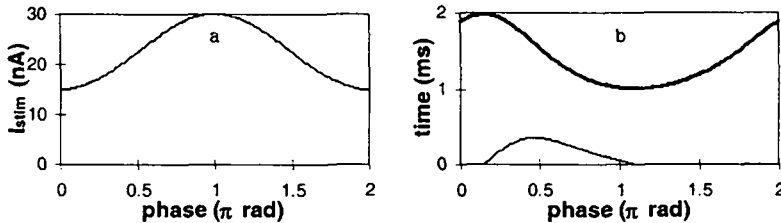
Figure 7.4 a) Stimulus current, and b) T_I (bold) and ΔT_R. See text for details.

Doubling the input amplitude reduces the integration time by about a factor of two. The exact amount is however a function of ϕ_s, f_{mod} and md, and can vary

widely with these parameters. Synchronisation will move to choppers with a larger T_R when the intensity increases. In Figure 7.4b we can see that in the above situation ΔT_R has almost exactly the same form as in Figure 7.3, but allows only half the spread in T_R at its maximum. This indicates that with increasing amplitude, fewer choppers will synchronise when we assume that their refractory periods are distributed evenly. This result is supported by biological evidence, showing that synchronisation of chopper neurons in the cochlear nucleus decreases with increasing stimulus level (Frisina, et al., 1990). There is further evidence that coincidence-detecting neurons in the inferior colliculus detect fewer coincidences at their best modulation frequency when the stimulus intensity increases (Rees and Palmer, 1989).

Figure 7.5 shows how the maximum value of ΔT_R varies with intensity for two different values of md and f_{mod}.
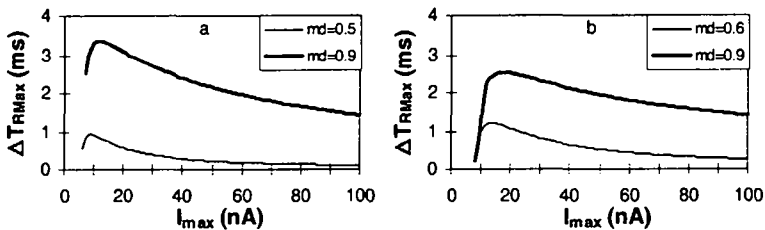
Figure 7.5 Maximum allowable spread in T_R for two different modulation depths and modulation frequencies. a) $f_{mod} = 100$ Hz or b) $f_{mod} = 150$ Hz.

The lower curve in Figure 7.5a is the theoretical result for the situations in Figure 7.3 and Figure 7.4. We can see that for $I_{max} > 10$ nA, ΔT_{RMax} reduces by about a factor of two for each doubling of I_{max}. However, when we increase the modulation depth (see the bold curve in Figure 7.5a), or the modulation frequency (see Figure 7.5b), ΔT_{RMax} varies less (in relative value) with I_{max}.
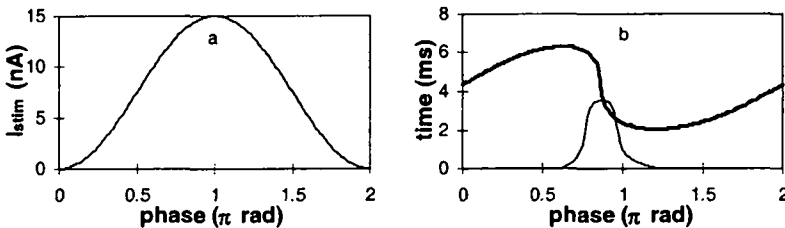

Figure 7.6 a) Stimulus current, and b) T_I (bold) and ΔT_R. See text for details.

When we use a larger modulation depth, the stimulus varies the integration time more; synchronisation will get stronger, allowing a much larger spread in

T_R and concentrating the spikes at a much smaller range of phases, as shown in the bold curves in Figure 7.5. We can also see this in Figure 7.6, where the modulation depth is 1.0.

The analysis above assumes that we have one spike per modulation period, and is only meant to clarify the synchronisation mechanism. For a complete description of the system, we would also have to analyse cases like two spikes per modulation period, or one spike per two modulation periods. Both cases can yield stable phase locking too, but are in general not as strong.

7.3 Implementation with the building blocks

I have used the building blocks described in chapters 3 and 5 to implement this envelope periodicity extraction model, as shown in Figure 7.7. The system of Figure 7.1 creates a two-dimensional tonotopic-periodotopic map, with one axis being position along the cochlea (i.e., the tonotopic axis which represents frequency on a logarithmic scale) and the second axis being modulation frequency. Taking a single output of the silicon cochlea, transforming it with the IHC circuit, and using this as input for the neuron chip allows us to model a single point in this map.
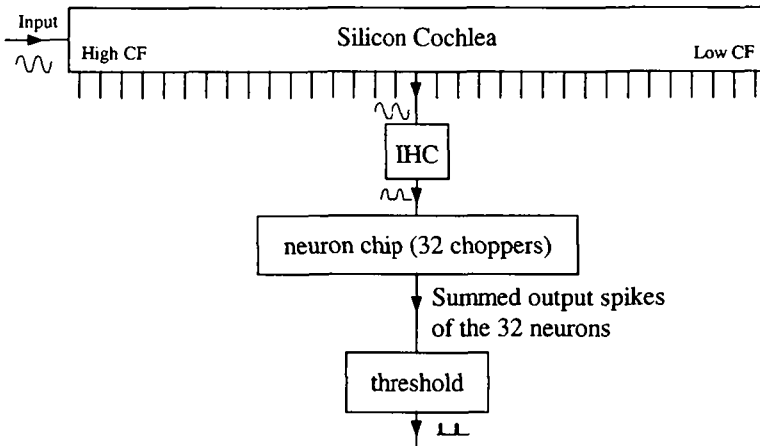
Figure 7.7 Implementation of the amplitude modulation sensitivity model with the building blocks.

The neuron chip can be set to simulate 32 similar sustained chopper cells by using a relatively high threshold voltage and low membrane leakage current, yielding a long integration time. The refractory period T_R can be controlled using the current I_{Kdown} (Figure 5.1). T_R will vary slightly among the neurons on

a chip, however, due to component mismatch. This mismatch will thus fix the ΔT_R in the group of choppers around a given T_R.

Since I have to rely on the statistical properties of the mismatch between the different neurons to create the spread in T_R, I shall need enough choppers in each group of similar choppers. For example, with only two choppers in a group there would be a high risk that their refractory period would be too different due to the device mismatch, and no synchronisation would be possible. The 32 neurons on a neuron chip represent a statistically large enough sample.

Synchronisation of the choppers can be detected by a coincidence-detecting neuron on a second chip, when the chopper output is used as input signal for this second chip. We can create coincidence-detecting cells on a second neuron chip by setting a relatively high leakage current, so that a number of spikes have to arrive within a certain time in order to evoke an action potential. The temporal window can be controlled by this leakage current.

A simpler solution to detect synchronicity of the chopper cells is, however, to look at the output line of the chip which contains the sum of all the chopper spikes at any given moment in time (see section 5.4.2). Each chopper injects a reference current on this wire when it spikes, so the currents of overlapping spikes will add. Synchronicity is detected by comparing the summed current with a threshold. For instance, using 16 times the reference current, at least half of the 32 neurons on the chip will have to have their spikes overlap to create a sum current larger than the comparator current. This solution does not allow for integration of non-overlapping spikes arriving in succession and the temporal window in which the spikes have to arrive in order to add up is controlled by the spike width. However it is a good enough solution to test the working principle of the model. I have used this solution in the measurements of the following section.

7.4 Test Results

7.4.1 Chopper response

The first step in the elaboration of the model is to test whether the group of spiking neurons on a chip acts like a group of similar choppers. We have already seen in the previous chapter that the neuron chip can recreate the Post Stimulus Time Histogram (PSTH) of the chopper. In Figure 7.8 two more examples of chopper PSTHs of the circuit are shown. They are the result of the summed response of the 32 neurons on chip to 20 repeated stimulations with a 5 kHz pure tone burst at two different intensities. These figures show that the response of the choppers yields a PSTH typical of chopping neurons, and that the chopping rate, keeping all other parameters constant, increases with increasing sound intensity. The chopping rate for an input signal of given

intensity can be controlled by setting the refractory period of the spiking neurons, and can thus be used to create the groups of choppers that differ in their average refractory period, shown in Figure 7.1. The chopping rate of the choppers in Figure 7.8 is about 300 Hz for a 29 dB input signal.
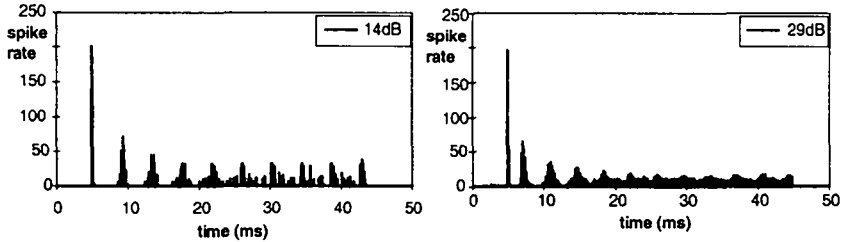

Figure 7.8 PSTH of the chopper chip for 2 different sound intensities

7.4.2 Coincidence-detection

Most neurophysiological data concerning low-frequency amplitude modulation of high-frequency carriers exists for carriers at about 5 kHz and a low modulation depth. I have therefore used a 5 kHz sinusoid in the following tests with a 50% modulation depth and stepped its modulation frequencies from 10 Hz to 550 Hz in steps of 10 Hz. Another free parameter of the model is the threshold current of the coincidence-detector. If this parameter is set so that at least 60% of the choppers must spike within the about 1 ms spike width to be considered a coincidence, we obtain the output of Figure 7.9. We can see that this yields the expected band-pass Modulation Transfer Function (MTF), and that the best modulation frequency for the 29 dB input signal corresponds to the intrinsic chopping rate of the group of neurons.
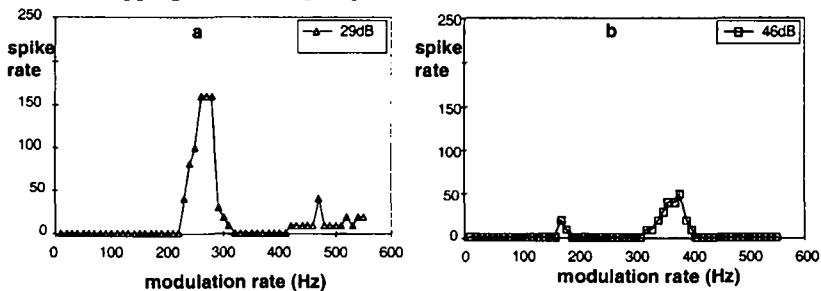
Figure 7.9 AM sensitivity of the coincidence-detector.

Figure 7.9 also shows that the best modulation frequency (BMF) increases with increasing sound intensity, just as the chopping rate in Figure 7.8. However, the maximum number of spikes per second actually decreases, which

is caused by the fact that ΔT_R is reduced when the input intensity is increased. In Figure 7.9b we can also see some evidence for the synchronisation with two spikes per period, as the choppers with a BMF of about 380 Hz also synchronise for a 170 Hz input. Note that second frequency is not necessarily half of the BMF. This is due to the fact that although T_R and T_S are the same for both spikes in the two spikes per period situation, T_I is not necessarily the same, since we integrate different sections of the waveform envelope.

7.4.3 *Influence of the coincidence threshold*

When the coincidence threshold is lowered to 50%, we can see in Figure 7.10a that the maximum number of spikes goes up, because the threshold is more easily reached. Furthermore, a second pass-band shows up at twice the best modulation frequency. At this higher frequency the choppers fire only every second amplitude modulation period; part of the group of choppers will synchronise during the odd periods, whereas the others will synchronise during the even periods. The division of the group of choppers will typically be close to, but hardly ever exactly 50-50, so that the 50% coincidence threshold is exceeded either during the odd or during the even modulation period. A 60% threshold will only rarely be exceeded, explaining the weak second peak seen around 500 Hz in Figure 7.9.
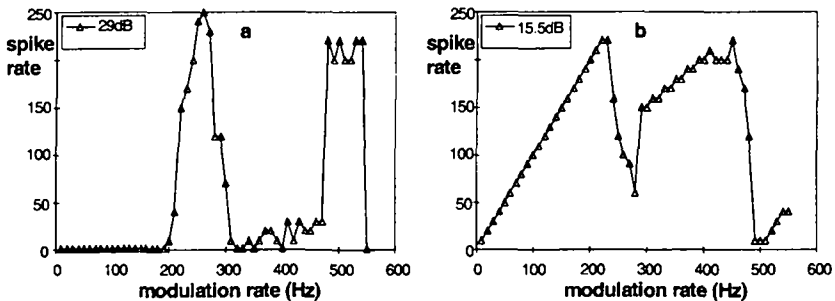

Figure 7.10 AM sensitivity of the coincidence-detector at lower threshold.

7.4.4 *Low intensity input*

Figure 7.10b shows the MTF for low-intensity signals with a 50% coincidence threshold. At low intensities the effect of the stimulation threshold shows up. Whenever the instantaneous value of the envelope is lower than the stimulation threshold, the spiking neuron will not be stimulated because its input current will be lower than the cell's leakage current. For low-intensity stimuli whose modulation frequency is lower than the group's average chopping frequency, the choppers will all come out of their refractory periods at a point in the modulation period during which the effective stimulation is zero (see Figure

7.11). They will therefore have to wait for the envelope amplitude to increase above the stimulation threshold before receiving stimulation anew. This waiting period nullifies the effect of the variation of the refractory period of the choppers, and thus synchronises the choppers for low modulation frequencies.
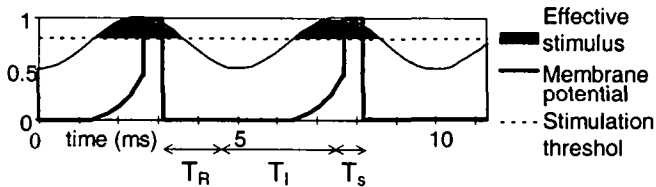

Figure 7.11 Spike generation for a chopper cell with stimulation threshold.

A second effect of this waiting period is that in this case the firing rate of the choppers matches the modulation frequency, since the choppers can only spike once per modulation period. This is equivalent to the observations that choppers in the cochlear nucleus have low-pass modulation transfer functions at low sound intensities, and band-pass modulation transfer functions at higher sound intensities (Frisina, 1990).

When the modulation frequency becomes higher than the maximum chopping frequency at low sound intensities, the choppers will fire only every second period, but will still be synchronised to the input waveform. All choppers do not necessarily synchronise to the same one out of every two cycles of the input waveform, but it is also unlikely that exactly 50% synchronises to one cycle and 50% to the other. Therefore the number of synchronised choppers will surpass the 50% threshold only once every two cycles, as can be seen between 300 Hz and 500 Hz in Figure 7.10b.

We can include this stimulation threshold nonlinearity introduced by the leakage current I_L in the mathematical analysis of section 7.2.2 by introducing a term $I_L T_I$ in equation (64) to obtain:

$$I_{max}\left(\left(1 - \frac{md}{2}\right)T_I - \frac{md(Sin(\phi_s) - Sin(\phi_s - 2\pi f_{mod}T_I))}{4\pi f_{mod}}\right) - I_L T_I - C_{mem}V_{thres} = 0 \quad (65)$$

From this we can see that increasing I_L has the same effect as decreasing the d.c. level of the stimulus current without changing the a.c. part.
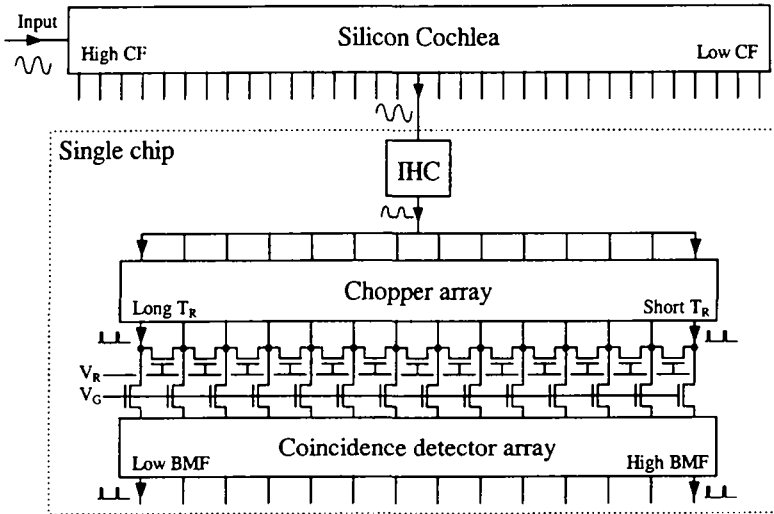
7.5 Second version

Figure 7.12 Implementation of the amplitude modulation sensitivity model.

I have recently implemented a new version containing the Inner Hair Cell model, an array of chopper cells, and an array of coincidence-detecting neurons on a single chip. However, instead of having several groups of similar chopper cells, as in the original model, this implementation uses an array of chopper cells along which the intrinsic chopping frequency increases from one end to the other.

Instead of connecting the output of a number of neighbouring choppers with dedicated wires to the input of each coincidence-detecting neuron, I have used the dendrite circuit of Figure 5.15 on this chip. As we have seen in chapter 5, this circuit offers a way to create a receptive field for the coincidence-detecting neuron, in which the chopper at the same array position receives the strongest input and the strength of the input falls off with distance.

7.6 Test results for the second version

7.6.1 Overview

A quick overview of the operation of this chip is shown in Figure 7.13 and Figure 7.14 which show measurements of the "cochlear nucleus" choppers and the "inferior colliculus" coincidence-detectors respectively.

Figure 7.13a shows the chopping rate of all choppers when stimulated with a 5 kHz pure tone. The chopping rate increases along the array for a given stimulus intensity, due to the decreasing refractory period of the choppers in the

array. Because of component mismatch, the chopping rate is quite a noisy function of the position in the array. Figure 7.13b shows the spiking rate of the choppers when stimulated with a 5 kHz signal which is amplitude modulated at 120 Hz at 50% modulation depth. We can see that the choppers around the 120 Hz position in the array (i.e., the choppers with intrinsic chopping frequency close to 120 Hz) fire once per modulation period, so that their spiking rates match the modulation frequency. This also means that such a chopper fires at the same phase at each modulation period, and it is therefore phase-locked to the modulation signal. The spiking rate for choppers having an intrinsic chopping rate much different from 120 Hz is however not changed by the presence of amplitude modulation.

Figure 7.14b shows the response of the IC coincidence-detecting neurons to the input from the CN choppers in Figure 7.13b. All choppers spiking at 120 spikes per second are synchronised to the input signal, but only a certain fraction of them spike at approximately the same phase in the modulation period. Only those choppers which spike within the time window of the coincidence-detectors will lead to a response from the coincidence-detectors. Figure 7.14a shows the response for a signal modulated at 130 Hz. The position of the responding coincidence-detectors moves to the right, i.e., towards the choppers with higher intrinsic chopping frequency.
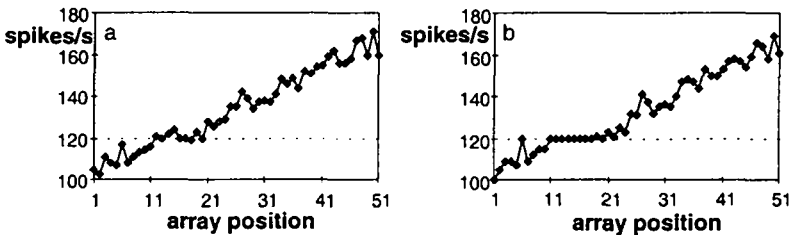

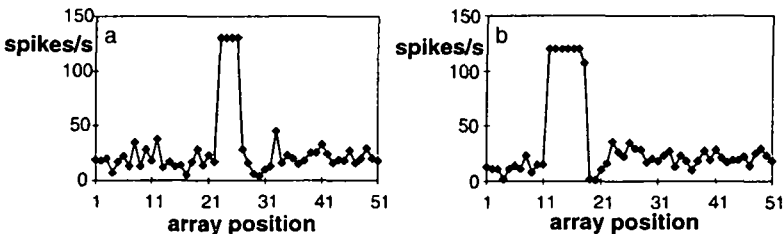**Figure 7.13** Average CN chopper spiking rate. a) no AM, b) 120 Hz AM at 50% depth.

Figure 7.14 AM sensitivity of the coincidence-detecting neuron. a) 130 Hz AM, b) 120 Hz AM.

We can also see that the maximum spiking rate of the coincidence-detectors matches the modulation frequency, since the choppers spike once per

modulation period and thus one coincidence is detected per modulation period. Therefore, the modulation frequency can be coded in both place and rate by these neurons. We will have to remember though that this circuit detects the modulation frequency based on one cochlear output only. In a complete system the modulation frequency would be represented by such a circuit at each cochlear output, and we will have to find a way to integrate the information across the tonotopic axis, i.e., over different chips. If modulation frequency is represented by a place code, tonotopic integration becomes quite trivial. It would be possible to simply sum the outputs of the coincidence-detectors with the same array position on the different chips. A rate code would necessitate the more complex operations of finding the rate that represents the modulation frequency for each chip, and comparing these rates between chips. A place code would therefore be our preferred way of coding the modulation frequency.

7.6.2 *Pure tone response*

After this overview of the operation of the second version of the model, we shall perform a more detailed exploration of this version in the rest of this chapter.
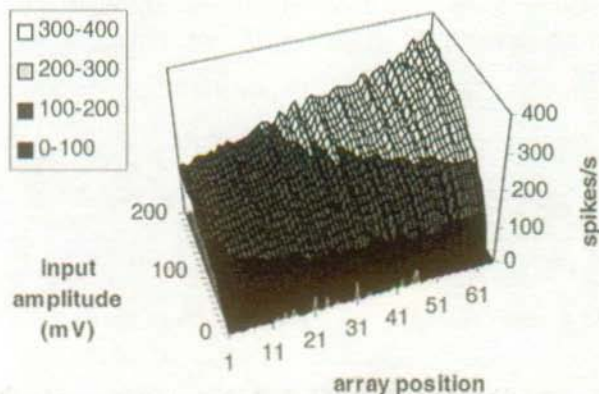

Figure 7.15 Spiking rate of the chopper neurons as a function of array position and input amplitude, when stimulated with a 3.6 kHz pure tone.

Figure 7.15 shows the average firing rate of the neurons in the chopper array of Figure 7.12 as a function of the input amplitude. The input signal was a pure tone at 3.6 kHz. In this and the following tests, this input signal has not been filtered by the cochlea, but stimulates the IHC circuit directly. The "input" signal should thus be seen as the output signal at a single output tap of the silicon cochlea. For a 3.6 kHz pure tone, the output of the IHC circuit simply becomes a d.c. current with only a very small remaining a.c. variation.

The output current of the IHC circuit is not equal to zero when the input amplitude is zero. This effect is equivalent to having a spontaneous rate on the auditory nerve. This stimulation current of the choppers has to be larger than their leakage current in order to elicit spikes. For the measurements of Figure 7.15 and the following tests, the leakage current was adapted so that it equals the amplitude of the stimulation current when the input voltage of the IHC circuit is zero. However, the input to each chopper is an imprecise copy of the stimulation current, and the leakage current of each chopper is also an imprecise copy of the bias leakage current. The zero input amplitude stimulation current will therefore be larger than I_{leak} for some choppers, and smaller for others. We can see this in Figure 7.15, since some choppers produce spikes even with zero input amplitude, whereas others don't.

Figure 7.15 also shows that the number of output spikes increases with increasing input amplitude, and that it increases faster for the choppers with a higher position number. Because of the shorter refractory period of the higher-numbered choppers, they will spike faster for the same stimulus level than the choppers with a longer refractory period. Furthermore, the maximum spiking rate of a chopper is limited by its refractory period, which explains the saturating nature of the spiking rate versus input amplitude curves of the choppers. At an input amplitude of 200 mV, we can see in Figure 7.15 that the spiking rate is still not completely saturated. This indicates that the integration time T_I is still a substantial part of the spiking period, since the integration time T_I is the only part of the spiking period that is influenced by the input amplitude — the refractory period T_R and the spike width T_S are constant. Since T_I is the only part influenced by the input signal, it is also responsible for chopper synchronisation by the input waveform. Therefore, we should not saturate the choppers if we want to be able to synchronise them.

The output current of the IHC circuit is also a saturating function of input amplitude, but this circuit has been adjusted to saturate at a higher input amplitude than the choppers, in order to allow the saturation related to the refractory period to show.

When high-frequency pure tones are used as inputs to the model, each chopper will chop at its own rate and no synchronisation will take place. The coincidence-detectors will thus only occasionally detect accidental coincidences.

7.6.3 Amplitude dependence

The second series of experiments show the response of the system to an amplitude-modulated 3.6 kHz pure tone as a function of input intensity for two different modulation depths (100% and 60%) and two different modulation frequencies (110 Hz and 150 Hz).

In the following figures, I have used a different representation for the output spikes of the coincidence-detecting neurons than for the output spikes of the choppers. In most of the following experiments either the choppers are inactive, or they spike at a rate close to the modulation frequency. This is clearly shown by an iso-intensity plot (Figure 7.16a) in which the lighter region represent active choppers, and the darker region represents the inactive choppers. However, the output of the choppers will only be partially modified from its basic form (shown in Figure 7.15) in the following experiments. An iso-intensity plot would not show this, so that I have opted to use the representation of Figure 7.15 for the chopper output. I cannot use this representation for the coincidence-detector output, because the active regions would hide the inactive regions that lie behind them in some cases.
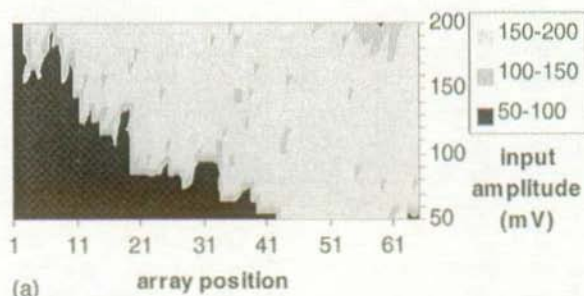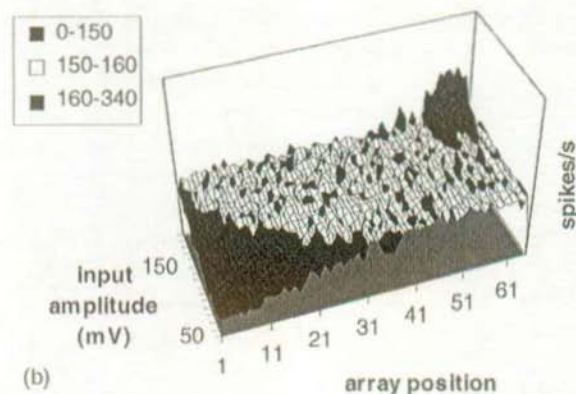
Figure 7.16 a) Spiking rate of the coincidence-detecting neurons, and b) of the choppers, when stimulated with a 3.6 kHz tone, 100% modulation depth, amplitude modulated at 150 Hz, as a function of maximum amplitude.

Figure 7.16 shows the result for a pure tone 100% amplitude modulated at 150 Hz. The 100% modulation ensures that a large number of choppers will be synchronised, shown by the large light plateau in Figure 7.16b. With increasing input intensity, the integration time T_I will shorten, and synchronisation will move to choppers with longer refractory periods T_R. Coincidences are only obtained once per modulation period when the choppers are synchronised. The active regions in the coincidence-detector output therefore correspond to the plateaux in the chopper output, but not all choppers that are synchronised to the input signal spike at the same time, and thus they do not always yield coincidences.
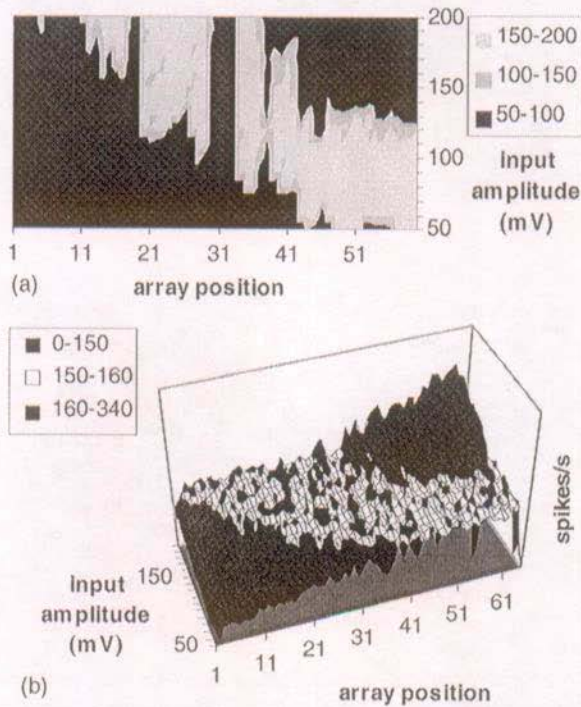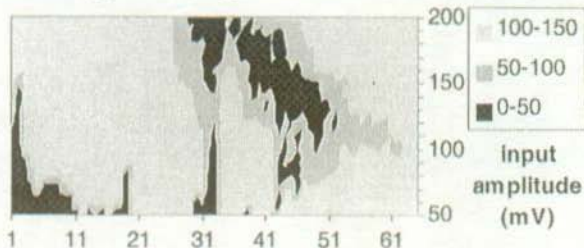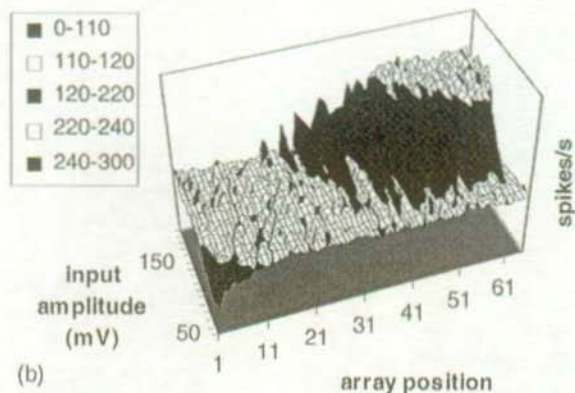

Figure 7.17 a) Spiking rate of the coincidence-detecting neurons, and b) of the choppers, when stimulated with a 3.6 kHz tone, 60% modulation depth, amplitude modulated at 150 Hz, as a function of maximum amplitude.

When we decrease the modulation depth, synchronisation will become less strong, and less choppers will synchronise. We can see this in Figure 7.17. Figure 7.17a also shows that even when choppers spike at the same frequency as the modulation frequency and are thus synchronised to the input signal, it

does not necessarily follow that neighbouring choppers are synchronised to each other. As we have seen in Figure 7.15, some choppers receive a higher input signal than others due to component mismatch. This means that these choppers will synchronise to different phases of the input signal, and thus will not create coincidences. A second reason for the fact that some coincidence-detectors do not always (or not at all in some cases) detect the synchronisation of neighbouring choppers is that these coincidence-detectors receive a weaker input signal, again due to component mismatch. These coincidence-detectors will thus need more choppers to be synchronised in order to spike.

Figure 7.18 a) Spiking rate of the coincidence-detecting neurons, and b) of the choppers, when stimulated with a 3.6 kHz tone, 100% modulation depth, amplitude modulated at 110 Hz, as a function of maximum amplitude.

Figure 7.18 shows the response with 100% amplitude modulation at 110 Hz. Again a large number of choppers are synchronised at this modulation depth, but at higher amplitudes the choppers with the shorter refractory periods will spike twice per modulation period. The second spike within the modulation period will occur on the falling flank of the envelope, and therefore does not receive the stabilising negative feedback on its integration time that the first

spike receives. The second spike will therefore not be synchronised with the second spike of the neighbouring chopper. In Figure 7.18a we can see that indeed the coincidence-detectors will detect only one coincidence per modulation period, even when the choppers spike twice per modulation period.
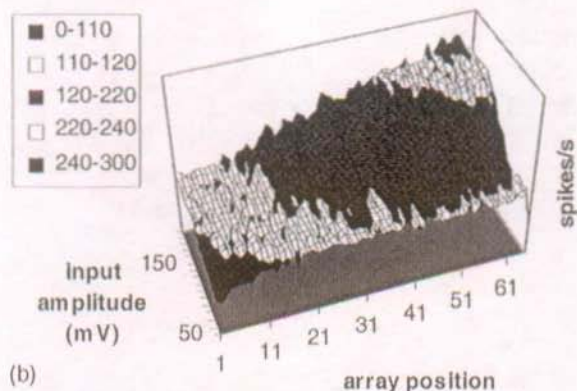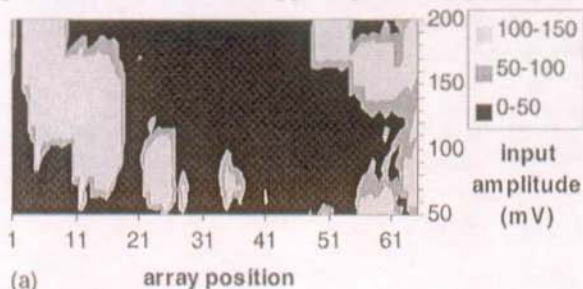

(a)

(b)

Figure 7.19 a) Spiking rate of the coincidence-detecting neurons, and b) of the choppers, when stimulated with a 3.6 kHz tone, 60% modulation depth, amplitude modulated at 110 Hz, as a function of maximum amplitude.

Finally Figure 7.19 shows that we also get fewer choppers that spike twice per modulation period and that have their first spike synchronised, when the modulation depth is decreased.

7.6.4 Modulation depth dependence

In the third series of experiments we will take a more detailed look at how the response of the system changes with modulation depth. In these experiments the maximum amplitude stays constant, and only the depth of the modulation changes.
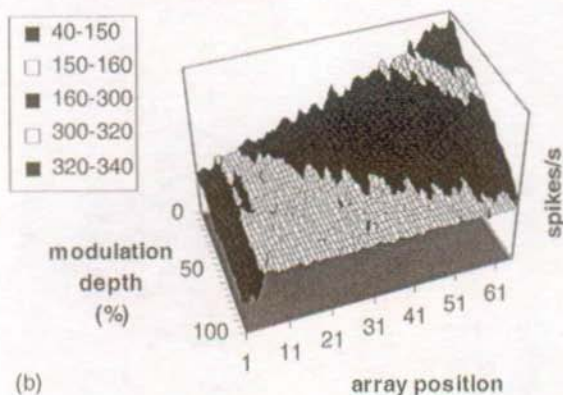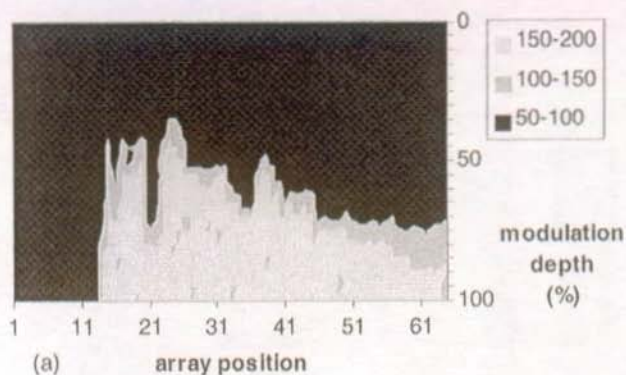
Figure 7.20 a) Spiking rate of the coincidence-detecting neurons, and b) of the choppers, when stimulated with a 3.6 kHz tone, 150 mV maximum amplitude, amplitude modulated at 150 Hz, as a function of modulation depth.

Figure 7.20 shows that the number of synchronised choppers decreases with decreasing modulation depth. However, the synchronised chopper that has the longest refractory period, i.e., the left-most synchronised chopper, remains the same independent of modulation depth.
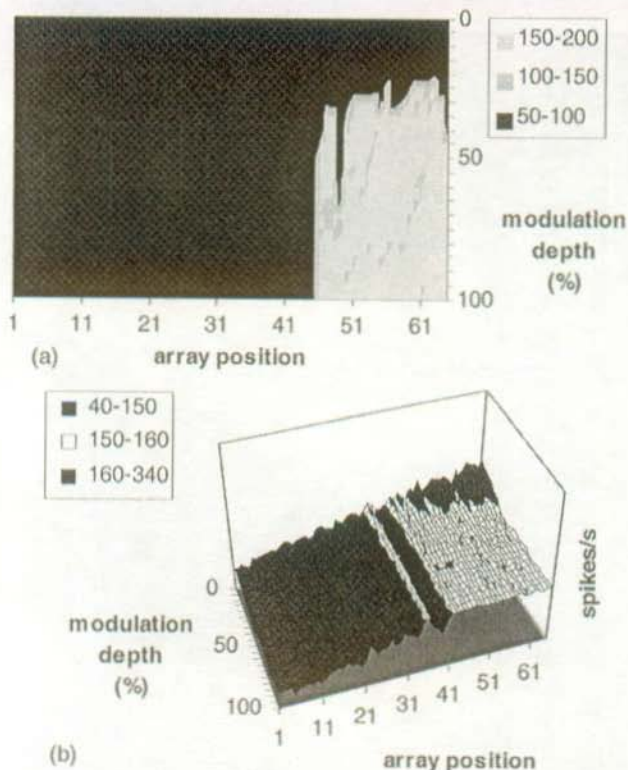
Figure 7.21 a) Spiking rate of the coincidence-detecting neurons, and b) of the choppers, when stimulated with a 3.6 kHz tone, 50 mV maximum amplitude, amplitude modulated at 150 Hz, as a function of modulation depth.

When we repeat the experiment at a lower input intensity, we obtain the results of Figure 7.21. Compared to the results in Figure 7.20, the group of synchronised choppers has moved to the right, i.e., to choppers with a shorter refractory period. This is as expected because the integration time will be longer at a lower amplitude. A second effect of reducing the input amplitude is that we obtain synchronised choppers at lower modulation depths than before. This conforms to the analysis in section 7.2.2 which showed that synchronisation gets stronger with decreasing amplitude at a given modulation depth.
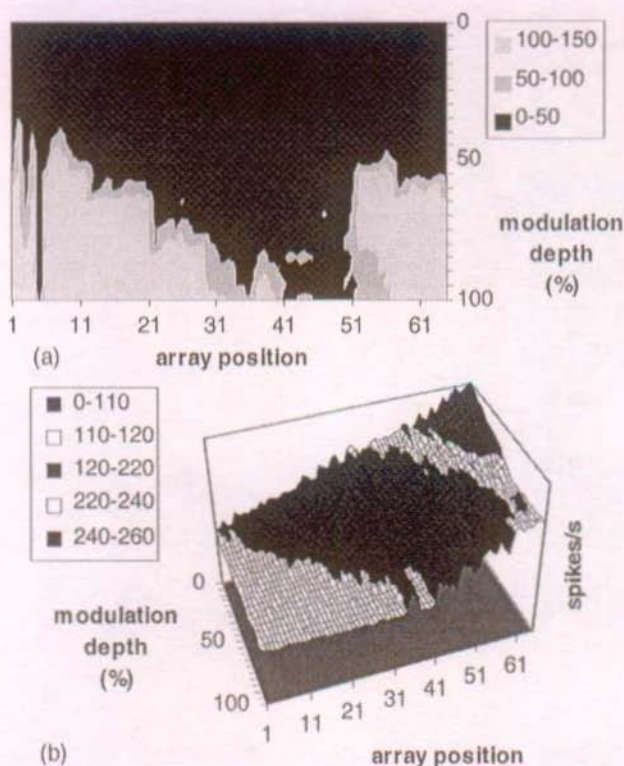
Figure 7.22 a) Spiking rate of the coincidence-detecting neurons, and b) of the choppers, when stimulated with a 3.6 kHz tone, 150 mV maximum amplitude, amplitude modulated at 110 Hz, as a function of modulation depth.

Finally Figure 7.22 shows the response at the higher input intensity (150 mV) to a modulation at 110 Hz. Again we obtain choppers that issue two spikes per modulation period, only the first of which is synchronised with the spikes of neighbouring choppers. Furthermore, synchronisation of the chopper with one spike per modulation period continues at lower modulation depth than synchronisation of choppers with two spikes per modulation period.

7.6.5 Stimulation threshold dependence

Up to now, the leakage current has been adjusted so that it is equal to the output current of the IHC circuit when the input to the IHC circuit is 0 V. Thus the effective stimulation current ($I_{stim} - I_{leak}$) of the choppers is zero when the envelope of the signal becomes zero, i.e., at the lowest point of the envelope when the modulation depth is 100%. We can also increase this leakage current so that the effective stimulation current becomes zero at a lower modulation

depth. Figure 7.23 shows the result for a 30% modulation depth at 110 Hz (a) and 150 Hz (b) when the leakage current is gradually increased. The coincidence-detectors that do not respond even when their input choppers are synchronised have been removed from these figures.
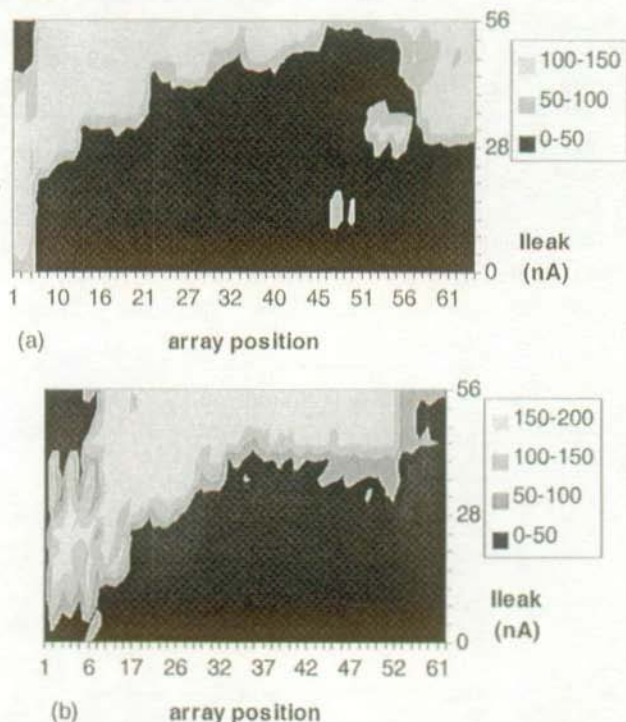

(a)   **array position**

(b)   **array position**

Figure 7.23 Spiking rate of the coincidence-detecting neurons when stimulated with a 3.6 kHz tone, 30% modulation depth, amplitude modulated at (a) 110 Hz, and (b) 150 Hz as a function of I_{leak}.

As predicted by equation (65) the number of synchronised choppers and the integration time T_I will both increase with increasing I_{leak}. Therefore, the first chopper in the array to be synchronised by the input signal will move to the right, i.e., it will be a chopper with a shorter refractory period. Furthermore we can again see a response generated by choppers that spike twice per modulation period for the 110 Hz modulation in the top right corner of Figure 7.23a.

7.6.6 Correction for amplitude dependence

We have seen that a different group of coincidence-detectors react to the same modulation frequency when the amplitude of the input signal changes. This poses a problem when we want to integrate information from different

chips connected to different cochlear outputs. The amplitude of the cochlear output signal will be different at the different cochlear output taps for any signal other than white noise. This means that we cannot simply sum the output of the coincidence-detector with the same array position at each cochlear output to obtain a collective estimate of the envelope periodicity.

A possible solution to this problem is to keep the average stimulation to the choppers constant, even when the cochlear output signal changes in amplitude. This can be very simply implemented by connecting the output of the cochlea to the input of the IHC circuit with a capacitor while keeping the average voltage input to the IHC circuit constant with a resistor, as shown in Figure 7.24. (For the significance of the IHC inputs V_{sig} and V_{ref}, please refer to Figure 3.15 in chapter 3.) This structure implements a high-pass filter with a corner frequency given by $RC/2\pi$. By choosing RC large enough we assure that the filter effectively functions as a d.c.-normalisation filter, giving all the signals the same d.c. level.
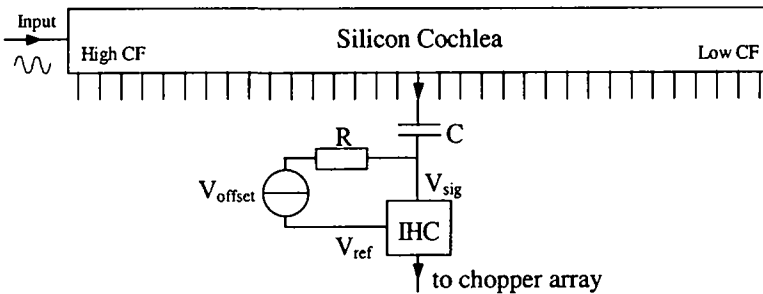

Figure 7.24 Implementation of the amplitude dependence correction.

This structure solves a second problem at the same time. Up to now, I have assumed that the average voltage of the cochlear output was always the same as the reference voltage of the IHC input. Both the reference voltage for the cochlear signals as for the IHC input can be controlled externally and can thus be set to the same value. The exact value of the reference voltage of the cochlear output signals will however vary due to component mismatch, thereby creating an offset input to the IHC circuit, which in turn creates a d.c. stimulation for the choppers. The structure of Figure 7.24 also solves this problem, since it restores the d.c. level of V_{sig} to $V_{ref} + V_{offset}$.
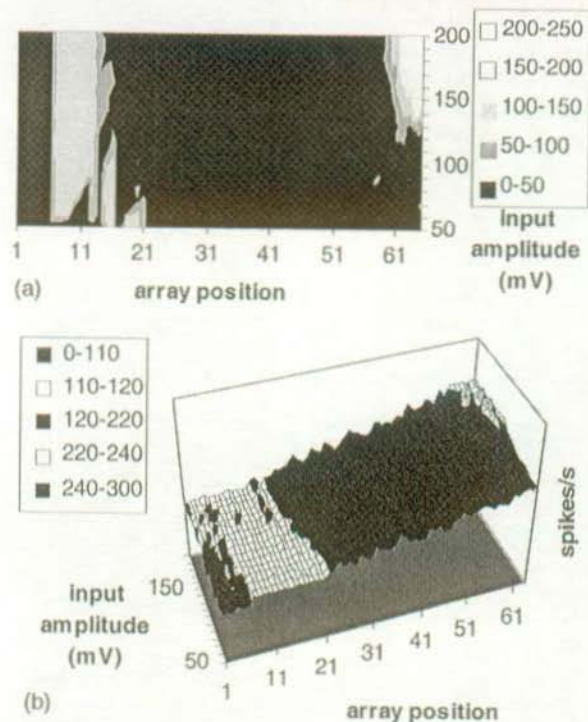
Figure 7.25 a) Spiking rate of the coincidence-detecting neurons, and b) of the choppers, when stimulated with a 3.6 kHz tone, 100% modulation depth, amplitude modulated at 110 Hz, as a function of maximum amplitude in the modified system.

Figure 7.25 shows the response of the modified system. We can see that in this case the group of choppers synchronised by the 110 Hz amplitude modulation is almost independent of input amplitude. Furthermore the addition of the voltage offset to the IHC input creates an offset stimulation current for the choppers that reduces the number of choppers synchronised by the input signal. An unexpected effect of this adjustment is the fact that the choppers that synchronise with two spikes per modulation period now have both spikes synchronised, so that the coincidence-detectors in the upper right corner of Figure 7.25a now spike at 220 Hz.
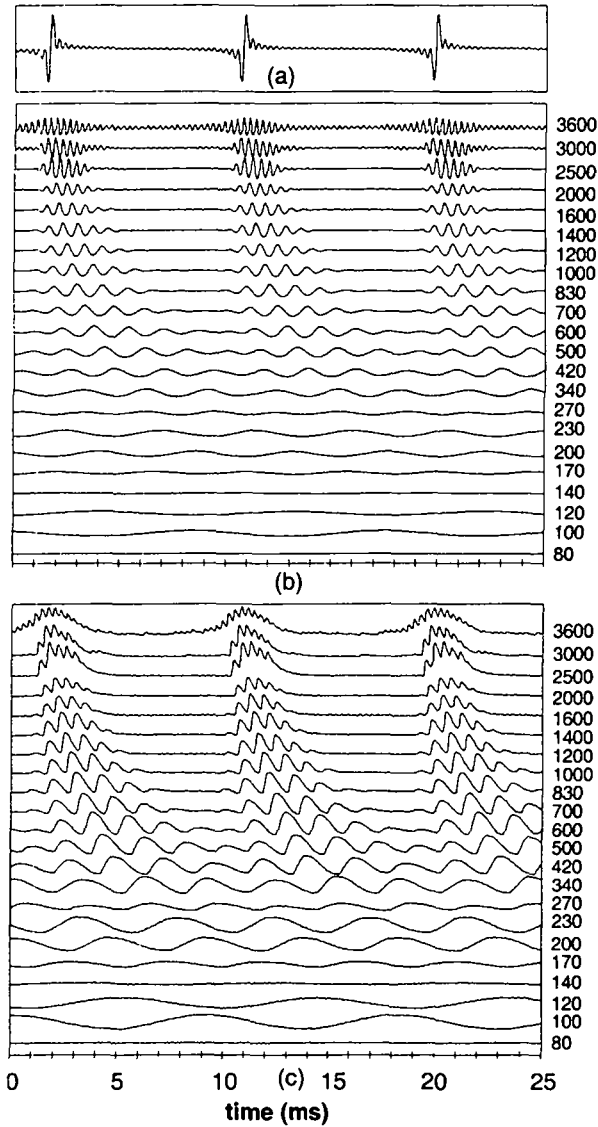
Figure 7.26 a) Input signal consisting of the first 32 harmonics of 110 Hz in sine phase at 60 mV input amplitude, b) measured output of the cochlear filters, and c) of the IHC circuit.

7.6.7 Sine phase harmonic complex response

To look at the response of the system at the different cochlear outputs, i.e., to create a complete tonotopic-periodotopic map, I have used a stimulus consisting of the first 32 harmonics of 110 Hz in sine phase, as shown in Figure 7.26a. This stimulus can be described by:

$$A \sum_{n=1}^{32} \text{Sin}(n2\pi f_0 t) \tag{66}$$

The output of the cochlear filters are shown in Figure 7.26b, and the output of the IHC circuit, (connected to the different cochlear outputs using the d.c. normalisation filter of the Figure 7.24) is shown in Figure 7.26c. The best frequency of the cochlear filters are indicated next to the traces.

The lower harmonics of the input signal will be resolved by the cochlear filter in the biological case and this has important psychoacoustical consequences, as we will see for instance in the next section. Due to the low selectivity of the silicon cochlear filters, the silicon cochlea only resolves the fundamental frequency. However, we can artificially augment the selectivity of the cochlear filters by pre-filtering the input signal to the cochlea with a band-pass filter adapted to the cochlear output that is being measured. I have done this for the measurements in this section and the next, using band-pass filters at the best frequencies of the cochlear outputs. Figure 7.26b thus actually shows the output of the silicon cochlea at the different output taps in response to different pre-filtered versions of the signal of Figure 7.26a.

Figure 7.26b shows that the first three harmonics of harmonics of the 110 Hz fundamental are resolved by the cochlear filters, so that the output of the cochlear filters with Best Frequencies (BFs) of 100 Hz, 120 Hz, 200 Hz, 230 Hz, and 340 Hz are sine waves at 110 Hz, 220 Hz and 330 Hz respectively. Furthermore, we can see that the filters with best frequencies in between these harmonics (80 Hz, 140 Hz, 170 Hz, and 270 Hz BF) hardly respond at all. For the filters with higher best frequencies, multiple harmonics are clearly present in the outputs. At first (420 Hz - 600 Hz BF) only two or three harmonics are present in the filters output, resulting in an output signal that is amplitude modulated with a modulation depth that increases as the best frequency of the filters increases. As the number of harmonics present in a filter's output increases, the time during which the output of the cochlear filter is zero increases (700 Hz - 2500 Hz BF). Finally, at the highest best frequency (3600 Hz BF) the number of harmonics in the filters output is reduced again, because the highest frequency available in the harmonic complex is lower than the best frequency of the filter.

In Figure 7.26c we can see that at the lower best frequencies, the output of the IHC circuit follows the output signal from the cochlea with some deformation due to the non-linear input function which responds stronger in one direction than in the other. At the high best frequencies (2500 Hz - 3600 Hz), however, the low-pass filtering of the IHC circuit is visible. This, together with the non-linear input function of the IHC creates a bump in the IHC output signal whenever the input signal's amplitude (Figure 7.26a) is high.
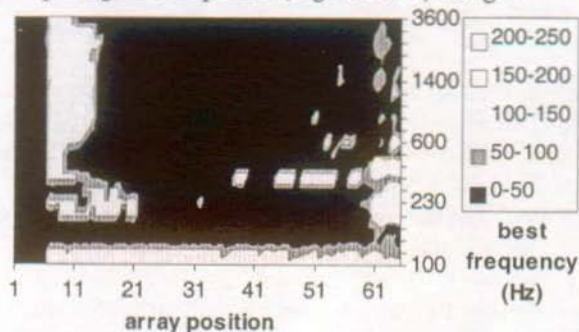

Figure 7.27 Spiking rate of the coincidence-detecting neurons when stimulated with the IHC outputs of Figure 7.26c.

Figure 7.27 shows the response of the coincidence-detecting neurons on the chip when the choppers are stimulated with the IHC signals of Figure 7.26c. In this figure the coincidence-detectors at position 8, 9, and 10 respond to the 110 Hz fundamental at the output of each cochlear filter, except at those cochlear filters that hardly respond to the input signal. At the output of the filters that resolve the fundamental frequency (100 Hz and 120 Hz BF), synchronisation is very strong and almost all choppers synchronise. Furthermore, the coincidence-detectors with a 220 Hz best modulation frequency, i.e. the coincidence detectors at positions 61 and higher in the array, also respond at the output of the cochlear filters with a best frequency close to the second and fourth harmonics of the input signal (200 Hz, 230 Hz, and 420 Hz BF). The same is true for the coincidence-detectors with a 165 Hz best modulation frequency (around position 51), which respond to every second period of the resolved 330 Hz third harmonic at the filter with a 340 Hz best frequency. At the output of the cochlear filters with a higher best frequency (from 500 Hz BF onwards) almost only the coincidence-detectors around position 11 respond and they do so at each bump in the IHC output signal.

Integration of the envelope periodicity information across the tonotopic axis, i.e., over the different cochlear outputs, seem quite straightforward in Figure 7.27. A summation in the vertical direction in this figure yields a clear peak at the positions 8 to 10, indicating the 110 Hz fundamental.
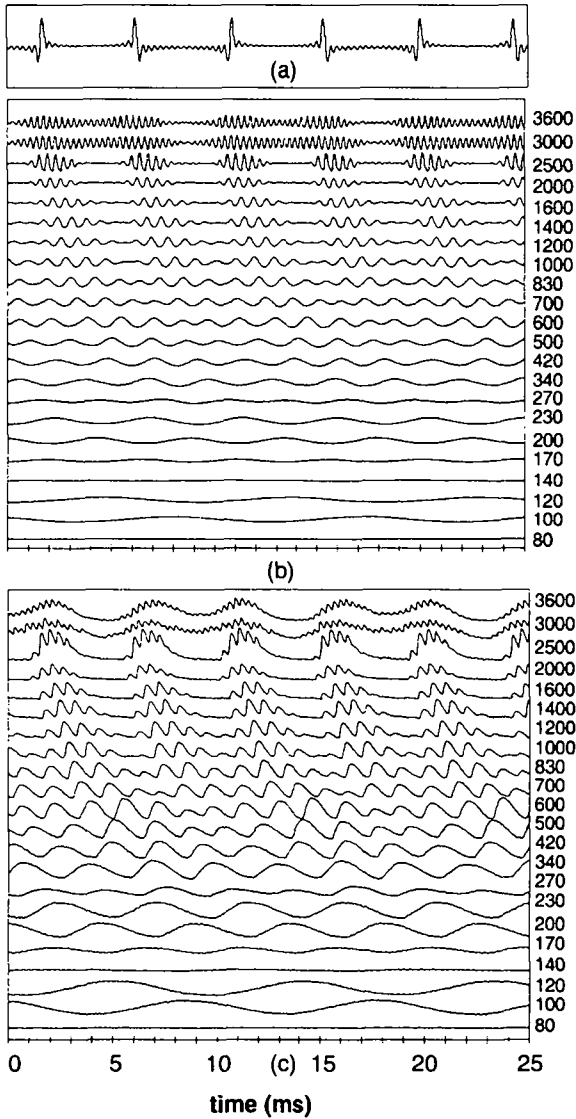
Figure 7.28 a) Input signal consisting of the first 32 harmonics of 110 Hz in alternating phase at 60 mV input amplitude, b) measured output of the cochlear filters, and c) of the IHC circuit.

7.6.8 Alternating phase harmonic complex response

The next experiment is performed with a harmonic complex for which the harmonics are in alternating phase (Figure 7.26a), described by:

$$A \sum_{n=1}^{16} \text{Sin}((2n-1)2\pi f_0 t) + \text{Cos}((2n)2\pi f_0 t) \qquad (67)$$

Psychoacoustical experiments have shown (Patterson, 1987) that when the complex consists of a number of the higher harmonics of the fundamental, the periodicity of a harmonic complex consisting of harmonics in alternating sine and cosine phase is judged twice as high by human listeners as the periodicity of the same complex but with all harmonics in the same phase. However, when the complex consists only of a number of lower harmonics, both in-phase and alternating-phase stimuli yield the same periodicity estimation. This is related to the fact that these lower harmonics are resolved by the cochlear filters, and thus do not interact with each other.

We can see in Figure 7.28b that for the cochlear filters that resolve individual components (100 Hz - 340 Hz BF), the periodicity of the signal envelope is not changed. Even for those filters whose output consists of two harmonics, for instance at 600 Hz best frequency, the envelope periodicity is not changed. However, for the filters with a high best frequency, the envelope periodicity is effectively doubled, just as the periodicity of the input signal of Figure 7.28a. Finally, in the middle range, at 700 Hz and 830 Hz best frequency, the 110 Hz periodicity is destroyed, while the 220 Hz periodicity is not yet present.

From Figure 7.28c we can see that the IHC output at the filters with a low best frequency as the same periodicity as the IHC output in Figure 7.26c. We thus expect that the coincidence-detectors respond in a similar way to both stimuli at the output of these low best frequency filters. However, at the higher best frequencies (from 830 Hz onward) the periodicity of the IHC output is effectively doubled in Figure 7.28c with respect to the IHC output in Figure 7.26c. At these best frequencies, we thus expect the coincidence-detectors to detect mainly the 220 Hz periodicity instead of the 110 Hz.
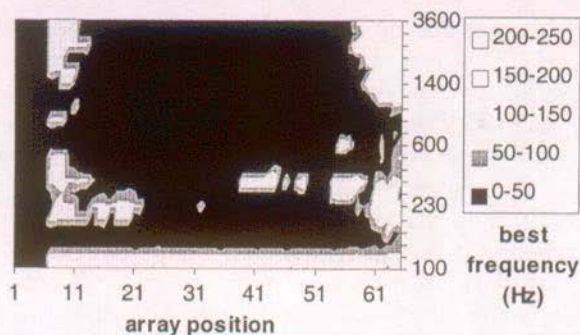
Figure 7.29 Spiking rate of the coincidence-detecting neurons when stimulated with the IHC outputs of Figure 7.28b.

Figure 7.29 shows the response of the coincidence-detectors to this input signal. We can see that compared to the output in Figure 7.27, nothing has changed indeed at the filters with a low best frequency, up to the filter with a 420 Hz best frequency. In the middle range, no periodicity is detected, because neither the 110 Hz, nor the 220 Hz periodicity is clearly present. Finally, at the high best frequencies, the periodicity is clearly dominated by the 220 Hz component as we can see from the strong response of the coincidence-detectors at positions 61 and higher at best frequencies of 1000 Hz and higher in Figure 7.29. Note that at these best frequencies the coincidence-detectors around position 11 also respond, signalling the presence of a 110 Hz periodicity. In fact, the choppers around position 11 are synchronised to one out of every two bumps in the IHC output of Figure 7.28c. The dominating response, however, is at 220 Hz best modulation frequency. The model thus predicts that the periodicity percept is the same for both in-phase and alternating phase stimuli when only the first four of five harmonics are present, while the periodicity percept will be doubled when we only present harmonics from the ninth harmonic on.

7.7 Summary

The neural system presented in this chapter uses the property of synchronisation of the chopper neurons to a particular amplitude modulation frequency. More generally, these choppers are synchronised by a small range of periodicities in the envelope of the signal. We can tune the choppers to be sensitive to different periodicities by changing their refractory period, and we can detect this synchronicity using coincidence-detectors. Using multiple coincidence-detectors with overlapping envelope periodicity sensitivity ranges, we can then extract the envelope periodicity of the signal

When an array of coincidence-detectors is connected to each output of the silicon cochlea, the output of the coincidence-detectors creates a map of signal frequency versus envelope periodicity. Using the silicon cochlea, the IHC chip, and the 32 neurons on the neuron chip as similar choppers, I have created the extraction for a single point in this tonotopic-periodotopic map by detecting coincidences of the output spikes of the choppers.

The integration of a dedicated chip allows the extraction of a range of envelope periodicities at a single cochlear output, thus substantially reducing the number of chips needed to implement a complete tonotopic-periodotopic map. This second version contains an IHC circuit, an array of 71 choppers with increasing chopping frequency along the array, and an array of 71 coincidence-detecting neurons.

The experiments with the sine phase harmonic complex and the alternating phase harmonic complex show two complete tonotopic-periodotopic maps. Furthermore, the alternating phase experiment effectively predicts the psycho-acoustic observation that the periodicity percept doubles for the higher harmonic numbers, but that the percept is unaffected at the lower harmonic numbers. We would need to repeat several more of these psycho-acoustic experiments to evaluate how well this system actually describes human pitch perception.

7.8 References for chapter 7

Frisina, R.D., Smith, R.L., and Chamberlain, S.C. (1990) "Encoding of amplitude modulation in the gerbil cochlear nucleus: I. A hierachy of enhancement," *Hearing Research*, Vol. 44, pp. 99-122.

Hewitt, M.J., Meddis, R., and Shackleton, T.M. (1992) "A computer model of a cochlear-nucleus stellate cell: Responses to amplitude-modulated and pure-tone stimuli," *Journal of the Acoustical Society of America*, Vol. 91, pp. 2096-2109.

Hewitt, M.J. and Meddis, R. (1994) "A computer model of amplitude-modulation sensivity of single units in the inferior colliculus." *Journal of the Acoustical Society of America*, Vol. 95, pp. 2145-2159.

Kim, D.O., Sirianni, J.G., and Chang, S.O. (1990) "Response of DCN-PVCN neurons and auditory nerve fivers in unanesthetized decerebrate cats to AM and pure tones: Analysis with autocorrelation / power-spectrum," *Hearing Research*, Vol. 45, pp 95-113.

Patterson, R.D. (1987) "A pulse ribbon model off monaural phase perception," *Journal of the Acoustical Society of America*, Vol. 82, pp. 1560-1586.

Rees, A. and Palmer A.R. (1989) "Neuronal responses to amplitude-modulated and pure-tone stimuli in the guinea pig inferior colliculus, and their

modification by broadband noise," *Journal of the Acoustical Society of America*, Vol. 85, pp. 1978-1994.

Webster, D.B., Popper, A.N., and Fay, R.R., Editors (1992) *The Mammalian Auditory Pathway: Neuroanatomy*, Springer-Verlag, New York.

.

8. Conclusions

This thesis describes the development, implementation and testing of several analogue VLSI building blocks, which may be used to create models of neural systems in the auditory pathway. Two such models have been realised and have been described in chapters 6 and 7.

To my knowledge it is the first time that an approach based on building blocks on separate chips has been taken and that these building blocks have been applied to modelling the auditory pathway, creating multi-chip neural systems. The work in this thesis shows that this approach is feasible and flexible. The neuron chip's biasing scheme makes it possible to simulate a variety of physiologically different neuron behaviours, such as choppers or coincidence detectors. Furthermore, the building blocks can be combined in different ways to realise different processing functions; only a few possibilities have been explored here.

Valuable lessons have been learned from the implementation of these building blocks:

The silicon cochlea that I have implemented is able to emulate the filtering of the basilar membrane in the human cochlea, with a similar distribution of best-frequencies and similar low- and high-frequency cut-off slopes. However, four fundamental problems have been identified for this cochlear filter cascade:

1. Each filter delays the signal by an amount that depends on the filters cut-off frequency; the number of filters that we reasonably can use to cover the auditory frequency range is therefore limited if the last filters are to react to the onset of an input signal within a few milliseconds.
2. Using a cascade of filters means that if a filter early in the cascade is defective, all filters following this filter will receive an incorrect input signal.
3. Each filter generates noise; since the output of a filter is the input of the next filter in the cascade, this noise is accumulated along the cascade.
4. A frequency selectivity similar to that of the biological cochlea can only be obtained with very high quality factors, which is impossible in the cascade design; the amplified noise would already saturate the transconductance amplifiers in the filter at a much lower quality factors.

The first two problems can be solved by going to a parallel filter structure which will also improve the third problem since the noise is not accumulated in a purely parallel structure. However, a purely parallel structure, i.e., one with no interaction between filters, is impractical because each filter would have to be of very high order to yield the steep high frequency cut-off slopes of auditory filters. Therefore, some coupling between the filters will be necessary, and we will need to take care that noise is not accumulated due to this coupling. The fourth problem might be solved by using filters that suppress the side-bands instead of using filters with a high gain around the best frequency. In any case, improving the silicon cochlea remains an active area of present and future research.

The inner hair cell circuit shows how we can capture the half-wave rectifying, low-pass and temporal adaptation behaviour of the real inner hair cell in a simple way. The low-pass filters used in this model are highly non-linear, which make them unusable when the input varies over an order of magnitude or more, but at a given sound intensity good agreement can be obtained.

The so-called "log-domain filters", which I've only recently discovered, should allow the development of linear low-pass filters in the current domain with a few transistors only; they seem a good option to improve the behaviour of the inner hair cell circuit with different sound intensities. These linear low-pass filters would also allow direct mapping on silicon of the differential equations of now traditional models, such as Meddis' hair cell model, or McGregor's neuron model. Although these models might be more linear than the biological cells, and thus more linear than a VLSI implementation needs to be, the implementation of these linear models will make it easier for other scientists to eventually use these silicon building blocks to implement their neural models in real time. Investigation of log-domain filters as neural building blocks is one of my future projects.

The spiking neuron circuit, with its leaky integration, threshold voltage, and refractory period, has been proven to be flexible enough to generate post stimulus time histograms similar to those of the major types of neurons in the ventral cochlear nucleus. Although the circuit might need to be extended to allow correct emulation of neurons with a more complex ion channel structure, the range of behaviours that the current circuit can simulate represents an encouraging start. Furthermore, interaction between the electronic neurons has been modelled by a range of circuits which include electronic synapses, dendrites, and axons.

Spike-based communication can also be used for interaction between chips. I have included such a communication scheme on the neuron chip, in which each neuron sends a unique address with a fixed number of ones and zeros on a common bus when it spikes. At the input side of the chip is a similar bus with a decoding circuit at the input of each neuron which checks for a particular address. When two addresses collide on the bus, we may either reject this input, or stimulate all neurons that have addresses which could have created this collision. With either option, noise in the communication channel will increase gradually with the number of collisions. I have furthermore provided the possibility to stimulate all neurons with the same input signal. Finally, an output signal which is the sum of current spikes from all the neurons on the chip is also included.

Apart from the ease of sending the signal from one point to the other in neural tissue or between chips, the advantage of spike based computation seems to lie in the ease with which synchronicity between two signals or more can be extracted using coincidence detectors. Two models that extract the periodicity of a signal, i.e., the repetition rate of the peaks of the amplitude in a signal, have been implemented with the electronic building blocks and both use synchronisation and coincidence detection in order to do so.

The first model extracts periodicities of an input sound with a high selectivity, i.e., each output only responds to a very limited range of periodicities. The advantage over using linear highly-selective band-pass filters is that the spike-based implementation can be highly selective without the slow reaction time associated with high selectivity in linear systems. Furthermore, the output of the system is independent of the input intensity of the signal, which the output of the cochlea itself is not. This model is an example of what an engineer might do when asked to design a system that extracts the periodicity, independent of intensity, with a high selectivity but a fast response and we can clearly identify the advantages of this implementation.

The second system is a model of the sensitivity of neurons in the cochlear nucleus and inferior colliculus to the envelope periodicity of amplitude modulated sounds or harmonic complexes. This system has a much stronger biological foundation and is an example of how we may model biology with the electronic building blocks. The engineering advantages of such a system, however, are much harder to discern; we shall need a better understanding of the purpose of the envelope periodicity extraction in the brain to see the advantages of this implementation over other possible systems.

Finally, an important engineering question remains: "Can the hardware modelling of neural systems be extended beyond simulating the results of scientific studies to produce genuinely useful technologies?" The models of periodicity extraction are indicative of future progress in this respect but, in reality, acceptable proof can only be based on working examples. It is therefore imperative to integrate and test hypothesised models of neural circuits in the brain, to increase our understanding of the functioning of the brain, and eventually to show the advantages of the neural approach for useful products in certain practical cases. Hopefully, the building blocks presented in this thesis will play an important role in this process.

Curriculum Vitae

Surname: van Schaik   **First names:** Floris André

Date of birth: 27 April 1967   **Nationality:** Dutch

Civil status: married

Addresses:

home		work:	MANTRA - DI - EPFL
	Av. d'Epenex 5		1015 Lausanne
	1022 Chavannes		Switzerland
	Switzerland	email:	vschaik@di.epfl.ch

Languages:
Dutch (mother tongue), English, French, German

Studies:
1985 - 1990 "Ingenieur Elektrotechniek" (MSc. EE equivalent), Twente University, Enschede, the Netherlands.

Professional Activities:
1989: Student project (4 months) at the IMS (Institut für Mikroelektronik Stuttgart), Stuttgart, Germany
1991 - 1993 and 1995: Staff scientist at CSEM (Centre Suisse d'Electronique et de Microtechnique), Neuchâtel, Switzerland.
1993 - 1997: Research assistant at the MANTRA Center for Neuromimetic Systems, Swiss Federal Institute of Technology, Lausanne, Switzerland
1996: Visiting scientist (4 months) at SEDAL (System Engineering and Design Automation Laboratory), Sydney University, Sydney, Australia.