Neuromorphic Cochlear Implant

Jan-Tore Marienborg

Department of Informatics
Faculty of Mathematics and Natural Sciences
University of Oslo

Thesis submitted for the degree of Dr. Scient.

May 2007

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo.
No. 607

ISSN 1501-7710

Cover: Inger Sandved Anfinsen.
Printed in Norway: AiT e-dit AS, Oslo, 2007.

Produced in co-operation with Unipub AS.
The thesis is produced by Unipub AS merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Unipub AS is owned by
The University Foundation for Student Life (SiO)

Contents

Preface

At childhood my curiosity awakened a strong interest for electricity and electronics as many of my funniest toys was electric. These gadgets was really puzzling as they possessed many really incomprehensible properties, like how my radio controlled car could be controlled over a fairly large distance without a wire, and it became of course necessary to examine the matter closer. This often led to opened and perhaps destroyed toys, but unfortunately an ultimate understating could not be achieved this way. Thus eventually, after some self study and basic education, I was able to begin studies at the University of Oslo in 1991. Although the direction of my studies was directed towards electronics all the time, the interests for other disciplines started to evolve relatively early. This did not make me to want to change discipline; instead it inspired me to find new areas where electronics could be utilized. When eventually medical electronics was touched upon during the studies, it became clear that electronics really can make the difference of life or death for some people, and not only be used in entertainment devices or in toys.

During my master-thesis work cochlear implants was introduced to me by my supervisor, and the theme attracted my interest long before a Ph. D. thesis became a viable alternative. But I really wanted to study these devices in depth, and this made the decision easy when I was offered a Ph. D. position at the University of Oslo. Looking back, I am really glad I took the opportunity, and I feel it has given me a solid basis for my position at Texas Instruments Norway (former Chipcon).

The time working with the degree has been educationally rewarding, but it has even given me a valuable insight into medical disciplines. In particular discussions with people designing implants, both at MedEl in Austria and Cochlear Americas in Denver was very rewarding. I am very grateful for the hospitality and cooperation I was given by both these companies.

At the end I want to thank my supervisor, Tor Sverre Lande, for organizing contact and meetings with the previously mentioned companies, for the guidance through my work, and for not loosing the belief in that I would finish one day. Then I would like to thank my fellow students and staff for making my studies memorable, and for giving me valuable input to my work. And finally my family, in particular my wife for support and for silent acceptance for all the late hours I spent working with the thesis.

Part I

Theory, design and measurements

Chapter 1

Introduction

A general trend for new ground-breaking products appears to be that they consist of highly integrated clever combinations of complex systems, permitting ground breaking functionality in miniaturized devices. These products often introduce electronics into new arenas solving problems which only a few years ago appeared unsolvable.

One common trend for most of these products is to cut development time by increasing the ratio of clocked digital versus analog computation. The benefits of this strategy are well known, but the high level of abstraction during design tends to create sub-optimal systems, particularly when power and space consumption are the most important design criterions.

This thesis explores the possibility of making a high performance, truly large power efficient analogue system. A suitable candidate is Cochlear Implants (CI) which in short is a hearing aid for people with severe hearing impairment, or which are completely deaf. The patient is given a varying degree of hearing by electric stimulation of the nerve-cells of the inner ear, see section 1.2 for short history and introduction.

These implants have been in commercial usage for about two decades, but almost all implants use a digital sound processor to precondition the nerve stimuli. One major obstacle is that this computation is rather expensive as it is done in real time for a number of parallel channels.

Fortunately there is much knowledge of the function of the human ear. This gives us the opportunity to be inspired by nature's solution to this complex problem to create an efficient electronic model of the system.

1.1 Rationale

The key to designing large analogue systems is to work at the right abstraction level. For some sub-systems it is unnecessary to utilize the full freedom of analog design, allowing us to simplify the design process by raising this level entering a semi digital domain. This frees resources that can be focused on the design issues that really matters for overall system performance. This simplification increases the chances of a successful design using reasonable amounts of resources.

Digital systems are basically quantized systems, where analogue signal levels are interpreted as quantized values. But it is important to remember that most digital systems are clocked, and that this creates a quantized time-domain too. This creates

a hard limit on maximum data throughput through a single data channel due to the finite clock frequency, and the number of defined logic levels (usually two). A true analogue system is un-quantized, both in time and value. The resolution or dynamic range of these systems is determined by noise, bandwidth and upper limit on signal levels. In theory digital systems can approach the data throughput of analogue systems if the numbers of quantization levels are increased until the resolution are limited by background noise, and the clock frequency are increased until bandwidth limits are reached. But this requires analogue design and simulation methodologies which are much to labor intensive for common digital designs. Instead, either time or value can be continuous while the other is quantized.

Most switched capacitor circuits are quantized in time but continuous in value, like the input stages of a SC-$\Sigma\Delta$ converter. Asynchronous digital circuits are continuous in time and quantized in value. And as shown later in this thesis, it is possible to make AD converters which are closely related to $\Sigma\Delta$ converters quantizing only the value, but not the time domain. This permits design of simple post-processing circuitry while maintaining much of the efficiency of an analogue channel.

The major challenges of cochlear implants are that a relative large number of parallel sound channels must be generated in real time. Still power consumption must be really low as these systems are battery powered, preferably using small batteries capable of storing only small amounts of energy. Usually the necessary sound processing is done with a low-power DSP (digital sound processor), which easily can be reprogrammed to suite the patients need, and possibly with new and improved sound processing algorithms. But this freedom comes at a cost. Even though they are low-power DSP's, they will consume a not insignificant amount of energy when they generate the cochlear implant output.

This thesis investigate the possibility of using analog sound processing to do initial sound conditioning, and then convert the analog signals into asynchronous pulse-domain in the final stage of the processing. This division of the sound processing appears to be very close to natures choice in implementation of hearing, which common evolution theory suggests is a fairly optimal solution.

1.2 Introduction to cochlear implants

This section gives a brief overview of the cochlear implant history, followed by an introduction to cochlear implant architectures.

1.2.1 History

For many years, it has been a desire to create artificial hearing for deaf people. The first known attempt with electrical stimulation was performed around 1790 by Alessandro Volta [50], who placed metal rods in his own ears and connected them to a 50-volt circuit, experiencing a jolt and hearing a noise "like a thick boiling soup".

The modern history of cochlear implants starts in the sixties. A team lead by William House did experiments with electrical stimulation on the cochlea of deaf people [10]. These test were performed with a five channel implant, but the conclusion of the tests was that the most natural sound was generated when all electrodes was stimulated by the

same signal. This test resulted in the development of a single channel cochlear implant (House/3M) in the early seventies. The implant consists of an external sound processor amplifying the original sound, then the sound is passed through a 340-2700 Hz band pass filter, and finally it is modulated using a 16kHz carrier signal. This signal is transmitted to the implant through an induction coil after it is amplified by an output amplifier, and it directly stimulates the implanted electrode without any demodulation [18]. Although the patients got a kind of sound awareness, it was far from the sounds perceived by natural hearing, and it gave very limited help to understand speech.

Later the idea splitting the sound signal over multiple electrodes to mimic the function of the human cochlea was tried with modest success in 1978. A few years later, in 1982, this idea was commercialized by Cochlear Cooperation with the Nucleus CI22 implant [7]. Still the implant performance was fairly modest, and it usually functioned as a lip-reading aid which in addition gave a general sense of environmental sounds.

In the 1990s the implants significantly improved with SPEAK, Continuous Interleaved Sampling (CIS) and related speech processing strategies [38]. These strategies pass the sound through an amplifier, then a bank of band pass-filters before the sound is rectified and utilized to modulate non-overlapping pulses for the patients' electrode array. By only stimulating one electrode at a time, the patients became able to understand speech over a phone-line without any visual contact with the speaker, and with fairly good perceived sound quality. CI moved on from being an aid for notion of crude environmental sounds and lip reading, to a real substitute for natural hearing.

1.2.2 Overview

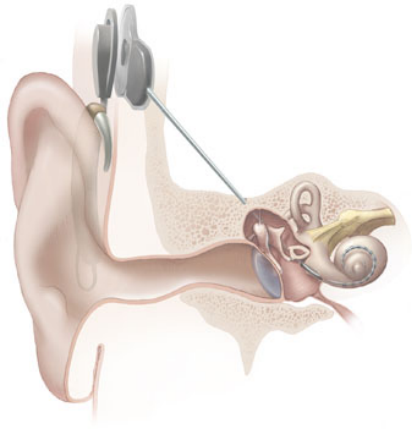


Figure 1.1: Illustration of a cochlear implant (Public domain, United States Department of Health and Human Services).

Most commercial cochlear implants are divided into an external and an internal part, see figure 1.1. Communication between these parts is done with an inductive coupling conveying both signal and power to the internal part. In the receiving end is

a coil with a magnet in its center, used to hold the transmitter coil into place behind the patients' ear. Close to the receiver coil is the receiver electronics, which again is connected to an electrode array floating in the cochlear duct. The receiver demodulates the signal and generates suitable electric signals for the electrode array. Due to high power-consumption, the sound processing and amplification is done externally, and only the necessary demodulation and pulse generation is done inside the body.

The two major components limiting the quality of the perceived sound are the sound processing algorithms and the electrode array. The sound processor replaces the function of a healthy cochlea, while the electrode array stimulates the auditory nerve-cells with the stimuli generated by the processor. There are currently from 6 to 22 electrodes on most implanted arrays, which are about two orders less than number of rows of hair-cells along the cochlear duct. This reduction in precision is most likely reducing the perceived sound quality, but unfortunately with the current technology, little gain is achieved when the number of electrodes is increased due to low precision in the neuron coupling. It is expected that ongoing research will manage to improve the electrode accuracy, permitting a larger number of stimulation channels.

The complexity and accuracy of the sound-processing algorithms are limited by current consumption and the accuracy of the electrode array. With a very limited number of channels, the processing algorithm that gives the best performance is not necessarily the one that copies the function of the cochlea most precisely. Thus if the number of channels is increased, it is likely that an extra bonus could be achieved if an algorithm more faithful to the architecture of the biological cochlea is implemented.

All together this will increase power-consumption and demands of the signal processor if the current sound processing strategies are modified to support the increased number of channels. This motivates the design of a new sound-processing unit, designed specifically for the task even at the transistor-level.

1.3 Premises

To make a significantly improved device, a number of premises have been identified. These premises can be grouped into two categories, comfort and sound quality.

In the comfort-group fall premises mostly affecting how the device is experienced in daily use. Typical issues are size, robustness and battery life. Sound quality is as the name implies the ability of the device to translate the external sound into a code, making the functioning nerve-cells respond such that the sound perception is as natural as possible.

In the first group falls as earlier noted the size of the device. This seems to be particularly important for children, partly because it might be a barrier for participation in common physical demanding play, as a large external device tends to get in the way and get damaged. If possible, it is desirable to make a fully implanted device, possibly with the microphone right behind the ear-drum. To do this, the power-consumption must be reduced to a minimum to permit small rechargeable batteries to be used. These could be charged during night through an inductive coupling for power transfer.

Another comfort criterion is that the device must have reasonable short reconfiguration latency when switching between different configurations for adaptation to varying listening conditions. If the delay is long, the period the device is in some unknown state

will be long enough to make annoying and possible painful noise. This will be even more important if the device is fitted with an automatic profile selection, but this is out of the scope of this thesis.

In the sound quality group falls advances in the quality of sound perception, patient dependent device adaptation, and as noted above, support for profiles tailored for different environmental listening conditions.

Clinical results show that a large percentage of the implanted patients do not get even close to natural hearing. It is my belief that this would be improved if a better model of the cochlea is used to generate the nerve stimuli.

Adaptability of the device to the patient is an important feature already implemented on existing commercial devices. This feature ensures that the nerve-stimuli is adapted to the remaining nerve functionality of the patient, and must be implemented on the test-chip to enable any useful comparison to existing devices.

Finally the patient should be given the opportunity to select from a number of profiles tailored for different listening conditions, intended to give the best possible performance for different types of background noise.

1.4 Architecture

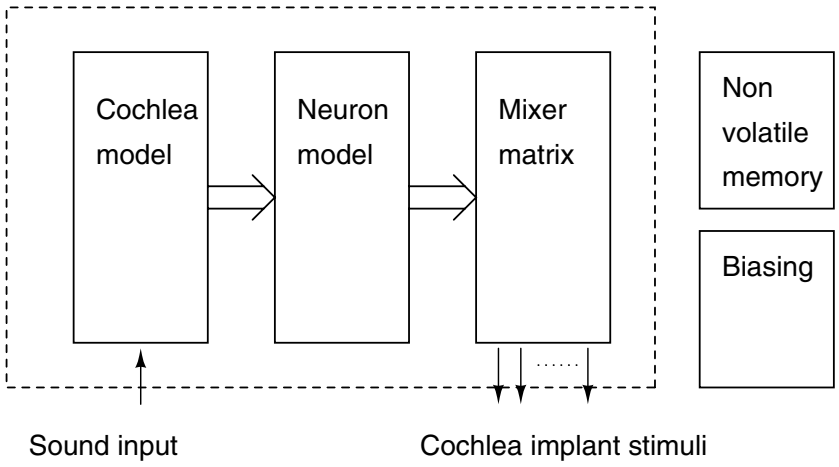


Figure 1.2: Block diagram of the design with blocks implemented on the test-chip in dashed box. Only the major signal flow is indicated.

The system consists of three main blocks which all have been implemented on a system test chip, see figure 1.2. The first block is the actual model of the human cochlea, which precondition the sound and splits it into a large number of analog components with different frequency content. This part of the circuit is based on an improved version of the model described in [41, 43]. The implemented cochlea is supporting a large number of channels, using only a single cascade to model the entire cochlea. This is in contradiction to the original design which needed three different sections to implement

the same number of outputs. The design and improvements are further elaborated in chapter 2.

The next block in the chain represents the neural system of the human cochlea. While these neurons would convert wave motion in fluid into nerve-pulses in a biological cochlea, they convert analog electrical signals into asynchronous voltage pulses in the electronic cochlea. This system is based on common implementations of neuron models, but they have been designed to be very power efficient and to accept voltage-mode inputs. An important detail improving the system performance is inhibitory feedback-networks, permitting the outputs of numerous neurons to be added to increase the signal-to-noise ratio of the resulting bit-stream beyond the capabilities of a single neuron. As explained in chapter 3 and chapter 4, there are a few studies showing that the axonal pulses of a biological neuron is very similar to the noise-shaped bit-stream of a $\Sigma\Delta$ converter, and that multiple bit-streams might be summed to improve the SNR beyond the performance of a single neuron with the aid of inhibitory feedback. This work demonstrates a novel implementation in analogue microelectronics of such a network in demonstrating that they are well suited for the generation of stimuli for cochlear implants.

The third and final major system building block is a signal mixer, necessary to condition the output of the nerve-cell models into the desired intensity and frequency bands of the implanted electrode array in the patient's ear. Chapter 4 gives a thorough description of this block. This signal mixing is to my knowledge novel, and in conjunction with the neuron bank this represents a new philosophy behind implementation of sound processing units.

As mentioned above, one test chip including necessary electronics for patient testing has been implemented, in addition some early test structures was implemented on a master-thesis test chip. As described later, the test-chip contains a minor error making patient testing difficult. Both chips are processed using a 1.2μm ORBIT process.

Chapter 2

Cochlea

2.1 Cochlear implants and the ear

Commercial available implants have matured to a level where the vast majority of users are able to understand speech without lip reading, generally making it easier for them to live in a modern society designed for hearing people. But still the sound is normally perceived as artificial with a metallic quality, the most common processor units are fairly large and the power consumption is high. All this reminds the wearer that he is hearing impaired, either if it is a child tangling up the wires and damaging this strange device or if it is a grown up which unfortunately forgotten spare batteries and suddenly lost the hearing in an awkward situation.

A major key to reduce the drawbacks of the implant and to enhance the general performance is to use a simple and robust signal processing system. The first stage of this processing is taken care of by an analog model of the human cochlea. The purpose of this model is to split the sound into a number of bands. Each band convey information of a narrow frequency component of the input signal, thus the nervous system is given specific information about the frequency content or timing of the input signal. The bands are finally encoded by the hair cells into nerve pulses when a portion of the basilar membrane vibrates at the local resonant frequency and bends the hair (Stereocilia) of the nearby hair cells. Thus the main challenges is to correctly model the pulse encoding, which is treated in chapter 4, and to model the function of the cochlea.

The movement of the stapes creates a traveling wave from base to apex in the scala vestibuli (see figure 2.1) through the opening in the oval window. The stiffness of the cochlear membrane is adapted to different resonant frequencies down the channel. It is relatively stiff at the base, creating resonant vibrations for high frequencies near the oval window, and both the stiffness and hence the resonant frequencies are decreasing towards the apex. These resonant vibrations of the membrane cause the hair cells in the organ of Corti, situated at the other side of the membrane, to generate nerve pulses. Thus the sound wave is split in a number of phase delayed frequency components creating a frequency to place transformation or tonotopic map of the input sound.

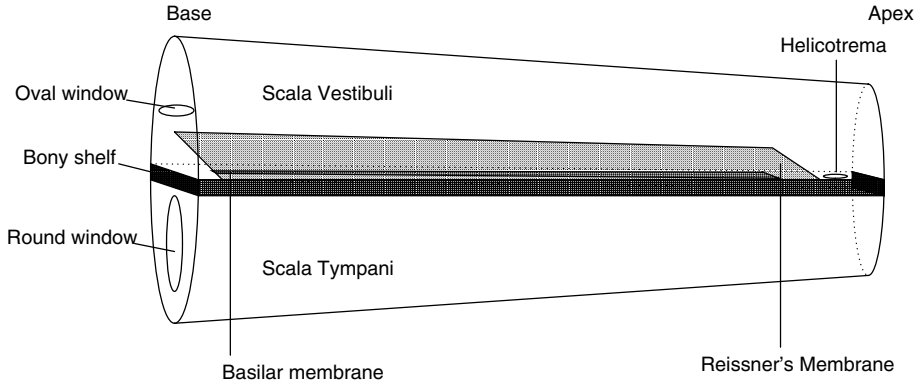


Figure 2.1: Stylized figure of the biological cochlea. The sound is transferred from the stapes in the middle ear to the fluid filled chamber scala vestibuli through the oval window. The wave travels in this medium from the base to the apex along the channel and back again to the round window in scala tympani through the helicotrema opening at the apex. In the length of the channel lies the basilar membrane which is stiff at the base, and gradually softening towards the apex, effectively gradually decreasing its resonant frequency down the channel. Below the basilar membrane are the spiral ganglion neurons which convert abrupt movement in the membrane to nerve pulses.

Timing information

Phase response or timing of the individual frequency bands is not widely acknowledged as an important information carrying parameter for speech processing, even though it is known that the cochlear duct creates distinct timing relationship of the different resonant places along the biological cochlea. Currently some implants can reverse the scan-direction of the output taps according to the patient´s desire. Since there is no clear evidence of a preferable scan-direction, it is usually concluded that the phase information in the signal is insignificant. But reversal of the scan-direction does not reverse the phase information on the modulated signal. Only by phase adjustment of the modulated frequency bands, this theory can be tested. To my knowledge no such experiments have been carried out yet with implanted patients.

From biology there are strong indications that the timing information is indeed significant. It is known that the percepted absolute timing of a sound transient onset is used in biaural cross correlation to locate the direction of a sound source. By correlation of the resulting neural responses an indication of the direction to the sound source is given. But even the phase relation of the different frequency bands is likely to carry important information.

When the wave travels down the cochlear duct the phase delay increases towards the low frequencies as the wave propagates from the base to apex. The typical delay is three to five cycles, which most likely is perceptible when interpreting transient sound like consonants or bangs. With no phase delay, the involved neurons will probably spike almost simultaneously, while the natural response creates an increasing response time from high to low frequencies spreading the pulses out in time. In chapter 4 it is

shown that the reduced probability of simultaneous pulses simplify the processing task for the implemented neuro-inspired circuit. There might even be a similar benefit for the biological system alleviating the neurons from massive simultaneous stimulation on the dendrites. The effect of this on the percepted sound is uncertain, but for the low frequency components of a transient the delay will probably be noticeable. Considering e.g. a 500Hz signal and a delay of five periods the total delay will be 10ms which probably is detectable.

Common commercial implants use filter banks which do not at all try to model the phase response of the biological cochlea. Thus the discussed effects will not be properly modeled, potentially distorting or removing useful sound information. But studies using implanted patients and the proposed implementation can unveil the importance of this effect.

2.1.1 Cochlear modeling

Three important methods for modeling of the biological cochlea with electronic circuits have been considered when selecting a preprocessor for the implant. The first approach utilizes a band-pass filter bank for division of the sound signal into different frequency bands. This approach is normally used in existing implant processors. Unfortunately this method does not try to match the phase response of a biological cochlea, possibly ignoring important information carrying parameters. More elaborate methods of cochlear modeling exist where the wave propagation is modeled also creating a more natural phase response. The first approach has been used in e.g. [31, 19, 8, 41] where the cochlea is modeled as a cascade of low-pass filters where the signal propagates through a series of slightly resonant filters with decreasing cutoff frequencies. Watts [49, 48, 47] approached the problem differently by modeling the wave medium of the cochlea as a two dimensional and bidirectional resistive network. To one edge of this network there are connected model circuits of the basilar membrane to complete the tonotopic map of decreasing resonant frequencies from base to apex. One major difference of this network compared to the cascade filter is that the model circuits are bidirectional, permitting wave reflections similar to observed effects in the biological cochlea. However the thesis concluded that this effect was of little importance, but the resulting modeling is more accurate than with a cascaded filter.

It is important to realize that at this point a very accurate model of the cochlea most likely is superfluous as the remaining cochlear spiral ganglion cells of a deaf patient can only be stimulated very crudely. In the implanted cochlea the electrodes are distributed along a flexible wire placed in the scala tympani below the bony shelf, while the spiral ganglion cells reside above this structure. Thus the distance between the electrodes and the nerve cells is relatively large, severely affecting the precision of the stimulation. In addition, the fluid in the chamber is conducting in all directions. A consequence is that there currently is a low limit of the usable number of electrodes [13, 14], typically an order of two magnitudes in difference between the number of electrodes from the cochlear implant and the number of spiral ganglion cells in a healthy cochlea. The resulting stimulation precision will therefore in itself reduce the quality of the perceived sound to a level which is likely to mask all minor improvements in the response of the cochlear model. Basically, if both the frequency and phase response of the cochlear

model roughly follows the response of a biological cochlea, it is likely that the perceived sound will be improved compared to the result with existing implants.

Because the aim is to create a sufficiently accurate sound processing unit, not a model of the human cochlea in its finest details, the priority will be correct phase and frequency response. From the analysis of the filter cascade given in [19], it is evident that the silicon cochlea fills these criteria. But as pointed out, this method requires special attention to the phase response of the filter to match the natural response. When the frequency resolution of such a cascade is increased by adding more filter sections with closer cutoff frequencies, the total delay of the cascade easily increase beyond what is desirable.

Watts points out that the Liouville-Green approximation used when creating the cascaded filter cochlear model poorly matches the actual internal frequency and phase response. But still the frequency and phase response at the best place can be fairly well matched to the physical response, thus the response of the cochlea outputs can be fairly accurate even though the actual internal signal propagation through the cascade poorly matches the desired response.

Even though the two dimensional model by Watts is more accurate, it is likely that the increased precision gained with this model makes little or no difference when used as an implant. Limitations imposed by the implementation of the passive components might easily give the model a worse total performance than the achieved performance of the cascade method.

In a biological cochlea some gain adjustment is believed to be performed by the inner hair cells to enhance the individual frequency bands. This function is only briefly treated at the end of Watts thesis, while the subject is thoroughly treated and incorporated in the solution proposed by Sarpeshkar [41]. Thus even though Watts propose a theoretically more correct model of the biological cochlea, physical limitations make the cascaded filter section model more promising.

See even article [12] included in part II for a general motivation on using filter cascades as cochlear implants.

2.2 Design

Based on the discussion in the previous section, a cochlear cascade of the type described in [41] is selected as the prime candidate for signal processing of the cochlear implant. The input to the cochlea will be a conditioned analog sound signal from a microphone when the circuit is used in an implant. The cochlea will in turn generate 100 analog voltage domain outputs which are converted to pulse domain and mixed in the mixer matrix described in chapter 4.

A quick look at the specifications above indicates that the size of the analog system will be very large. This requires special attention to robustness, stability, power, and area consumption. The selected design handles to a certain extent some processing mismatch by automatic offset and gain adjustment circuits, which will significantly reduce and equalize mismatch effects between different filter sections.

The area and power consumption is gracefully handled with sub-threshold filter sections, reducing the need for large capacitances. It has low W/L transistor ratios, and permits usage of the same basic filter sections for the complete filter cascade even though

the frequency range of the cutoff frequency is spanning three orders of magnitude. Even stability is nicely handled by the architecture, as discussed later, using a large number of similar and stable filter sections. These filter-sections constitute a stable filter of very high order capable of dividing the input signal into a large number of sub-bands even though the total area consumption is relatively small.

2.3 Cochlear implementation

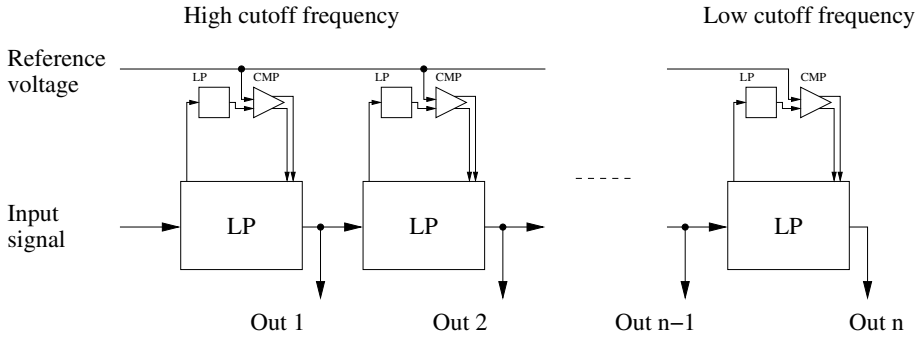


Figure 2.2: Block diagram of the cascade filter with offset correction circuit.

The design is entirely based upon the second order filter sections described in section 2.4 and the offset adaptation circuit described in section 2.6. While the original design by Sarpeshkar in [41] even includes the earlier mentioned inner hair cell model to obtain automatic gain adjustment, this feature has not been implemented on the test chip. The feature demands bipolar transistors and was originally planned to be implemented. The fabrication house indicated that a bipolar option would be available for the selected process during the second year of this work, but alas, this option never became a reality for MPW runs, and the circuitry proved very difficult to convert to plain MOS technology. Since this is not a part of the primary function to be tested in this thesis, it was finally decided to be left out.

In figure 2.2 a block diagram of the cascade is shown. To the left the signal is feed into the cascade, which corresponds to the base end of a biological cochlea. Then the wave propagates down the cascade through the filters with gradually lowered resonant frequencies to model the softening of the basilar membrane and the consecutive decreasing resonant frequency from the base to the apex, see figure 2.1. A logarithmic decreasing resonant frequency is implemented by connection of the second order filter section (SOS) V_{b1} and V_{b2} biases to serially coupled poly resistors constructing a multi output voltage divider. The external bias voltages are connected to both ends of the poly lines, creating a monotonically falling voltage gradient controlling the cutoff frequency of the sub-threshold SOS blocks, see figure 2.3. Since the bias voltages are connected to MOS transistor gates and the divider only depends on passive resistors, very low current is necessary through the voltage divider to generate very accurate and stable voltages. With the exponential g_m dependency of the operational amplifiers in the SOS blocks, the resonant frequency is exponential varying through the cascade. The

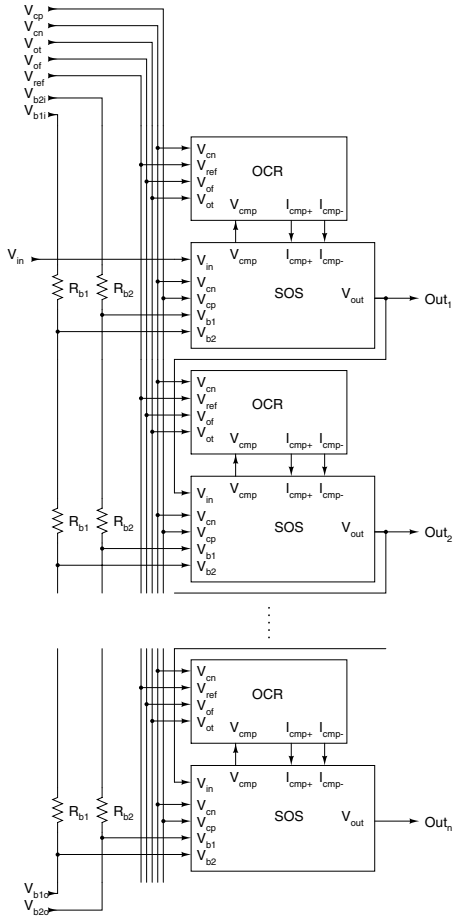


Figure 2.3: Schematic of the cochlea showing both the second order filter section (SOS) and offset correction circuitry (OCR). n=100 in the implemented cochlea.

resonant frequency is adjusted with the product of the bias voltages, while the Q-factor is adjusted with the ratio of the two bias voltages according to equation 2.9 and 2.10.

Filter stability is particularly important when the resonant peak of a feedback filter structure is used for band-pass filtering. To achieve good signal to stop-band ratio the Q-factor must be high, and for an active amplifier this is usually achieved with a positive feedback of the pass-band frequency. A strong feedback creates a sharp filter response, but unfortunately even reduces the phase margin potentially making the filter unstable. But the cascaded filter structure achieves high cumulative resonant peaks by partial overlap of the resonant frequencies. Thus high pseudo-resonant peaks and high effective Q is achieved using stable filter sections with very low Q. Since the filter cascade is composed of stable filter sections without global feedback, the complete cascade must be stable.

Figure 2.3 show the complete block level schematic of the cochlea. All bias voltages and currents are described in detail in section 2.4 and 2.6. To simplify the adjustment of the cascade, V_{rf} can be extracted from the input signal as described in [27], and hence removed. Further, I_{1i}, I_{1o}, I_{2i} and I_{2o} can be converted to voltages directly controlling the undamped resonant frequencies and the Q-factor. All the resulting and remaining bias voltages are constant during chip operation and can be set to constant values by on-chip bias circuitry on a final chip.

2.4 Filter section implementation

The main building block of the cochlear cascade is the second order sections. In the following chapter a thorough analysis of the section is performed to create a fundament for analysis and design of a complete cochlear cascade. To demonstrate the feasibility of the theory a description of the final design with selected simulation and measured results is included.

When implementing low frequency filters in VLSI technology, the size of the passive components are a problem if a linear RC-filter design strategy is used. One common solution to this problem is to use switched capacitor implementations where nonlinearity is exploited to achieve compact design despite the low cutoff frequency. But the large number of converters necessary to achieve the desired performance would probably both be space and power consuming in addition to creating troublesome switch noise. Instead the proposed design relies on a low power transconductance-C filter section utilizing the transconductance of the amplifier as a resistive element.

2.4.1 Linear analysis

The analysis of the second order section from [41] has been reviewed to establish a better design fundament for a cochlear cascade. Particularly interesting is the damping properties in the pass band, which is known to be challenging, and the variation of the cutoff frequency and quality factor in respect to g_m of the amplifiers. In the original evaluation in [41], these effects are only partially addressed, and with two different sets of equations. This section fills an earlier gap in the analysis of the design by introduction of equations for damping, Q-factor and cutoff frequency, and a discussion of their properties.

The transconductance-C [46] filter sections replace the output serial resistance with the transconductance (g_m) of the transconductance amplifiers (OTA), both permitting an easy adjustable cutoff frequency by g_m tuning, and a compact design. The drawback is that the transconductance-C filter [46] is sensitive to parasitic effects. To reduce power and space consumption, small capacitive load is desirable, thus requiring small g_m. But this makes parasitic capacitances significant compared to the intentionally implemented capacitances, and even the output transconductance (g_{mo}) becomes important compared to the transconductance.

The full schematics of the section is shown in figure 2.4, showing the filter section core consisting of the two WLR amplifiers and capacitors. The V_{cmp} and I_{cmp} terminals at the top of the schematic is for offset correction, which is further described in section 2.6.

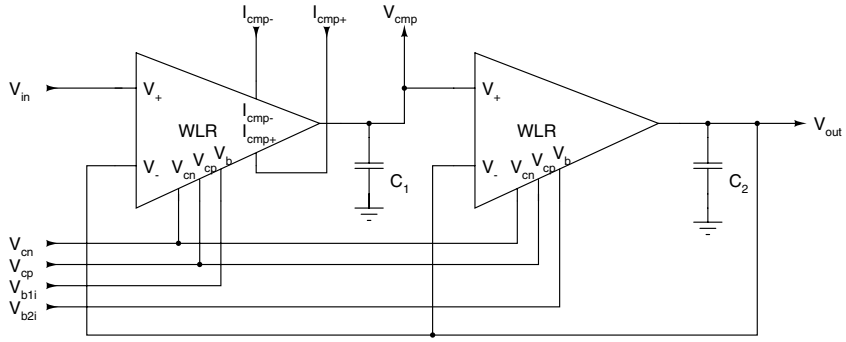


Figure 2.4: Block diagram of the full second order section.

Transfer function

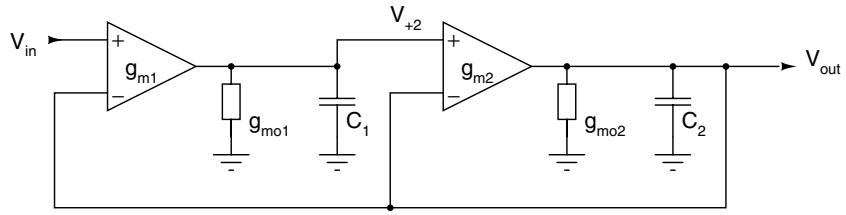


Figure 2.5: Model of the SOS-section with the output conductance separated as g_{mo1} and g_{mo2}.

The transfer function is found from the partly idealized model in figure 2.5. This model does not explicitly take into account the input capacitances of the transconductance amplifiers, because most of it can be included in C_1 and C_2. The parallel output amplifier conductance is modeled as shunt resistors because the circuit uses these actively instead of external serial impedance. The model can be transformed to the equivalent block-diagram in figure 2.6.

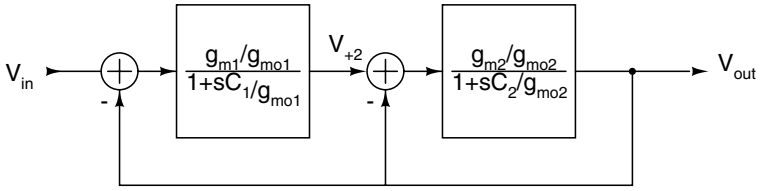


Figure 2.6: Block diagram of the SOS-section derived from figure 2.5.

g_{m1} and g_{m2} are the transconductance of the amplifiers while g_{mo1} and g_{mo2} are the output conductance of the amplifiers. Note that g_{mo1} usually equals g_{mo2}. Finally C_1 and C_2 are the parallel output capacitances including parasites.

It is shown in appendix A that:

$$H(s) = \frac{\alpha}{s^2\beta + s\gamma + 1} \tag{2.1}$$

where

$$\alpha = \frac{1}{\frac{g_{mo1}g_{mo2}}{g_{m1}g_{m2}} + \frac{g_{mo1}}{g_{m1}} + 1}$$
$$\beta = \frac{1}{\omega_1\omega_2(\frac{g_{m1}g_{m2}}{g_{mo1}g_{mo2}} + \frac{g_{m2}}{g_{mo2}} + 1)}$$
$$\gamma = (\frac{1}{\omega_1} + \frac{1}{\omega_2})\frac{1}{\frac{g_{m1}g_{m2}}{g_{mo1}g_{mo2}} + \frac{g_{m2}}{g_{mo2}} + 1} + \frac{1}{\omega_1}\frac{1}{\frac{g_{m1}}{g_{mo1}} + \frac{g_{mo2}}{g_{m2}} + 1}$$

In both part B and C the expression C_1/g_{mo1} is replaced with $1/\omega_1$ and C_2/g_{mo2} with $1/\omega_2$.

From equation 2.1 it can be seen that the filter will have damping in the pass-band if g_{m1} and g_{m2} are small, or g_{mo1} and g_{mo2} are large. Further g_{m1} and g_{mo1} affect the damping more than g_{m2} and g_{mo2}, and g_{m2} affect the cutoff-frequency more than g_{m1}.

Properties of the transfer function

The transfer function $H(s)$ can be written as:

$$H(s) = \frac{\delta}{(s/\omega_0)^2 + (s/\omega_0)(1/Q) + 1} \tag{2.2}$$

where δ is the damping factor, ω_0 is the undamped angular resonance frequency and Q is the quality factor.

The δ, ω_0 and Q-factor are easily read from $H(s)$ as:

$$\delta = \frac{1}{\frac{g_{mo1}g_{mo2}}{g_{m1}g_{m2}} + \frac{g_{mo1}}{g_{m1}} + 1} \tag{2.3}$$

$$\omega_0 = \sqrt{\omega_1\omega_2(\frac{g_{m1}g_{m2}}{g_{mo1}g_{mo2}} + \frac{g_{m2}}{g_{mo2}} + 1)} \tag{2.4}$$

$$Q = \frac{1}{\sqrt{\frac{\omega_1}{\omega_2}} + \sqrt{\frac{\omega_2}{\omega_1}}(1 + \frac{g_{m2}}{g_{mo2}})}\sqrt{\frac{g_{m1}g_{m2}}{g_{mo1}g_{mo2}} + \frac{g_{m2}}{g_{mo2}} + 1} \tag{2.5}$$

If $g_{mo1} = g_{mo2} = g_{mo}$ the equations reduce to:

$$\delta = \frac{1}{\frac{g_{mo}^2}{g_{m1}g_{m2}} + \frac{g_{mo}}{g_{m1}} + 1} \tag{2.6}$$

$$\omega_0 = \sqrt{\frac{g_{m1}g_{m2} + g_{m2}g_{mo} + g_{mo}^2}{C_1C_2}} \tag{2.7}$$

$$Q = \frac{1}{\sqrt{\frac{C_2}{C_1}} + \sqrt{\frac{C_1}{C_2}}(1 + \frac{g_{m2}}{g_{mo}})} \sqrt{\frac{g_{m1}g_{m2}}{g_{mo}^2} + \frac{g_{m2}}{g_{mo}} + 1} \tag{2.8}$$

In appendix A it is shown that the damping factor δ approaches unity if $\{g_{m1}, g_{m2}\} >> g_{mo}$. And it is shown that the equations for the angular resonance frequency and Q-factor can be further simplified if the same assumptions are made. Then the equations reduce to:

$$\omega_0 \approx \sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}} \tag{2.9}$$

$$Q \approx \sqrt{\frac{g_{m1}C_2}{g_{m2}C_1}} \tag{2.10}$$

Note that equation 2.9 and 2.10 equals the corresponding equations in [41] if $\tau_1 = C_1/g_{m1}$ and $\tau_2 = C_2/g_{m2}$

Comparison with the analysis from [41]

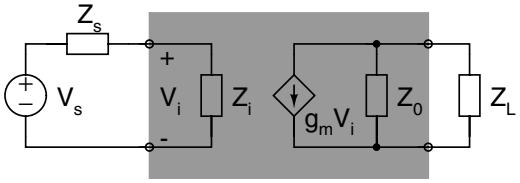


Figure 2.7: Electrical equivalent of the transconductance amplifier.

The results, particularly from equation 2.4 and 2.5 might be surprising compared to the analysis used in [41]. The differences stem from the fact that if the assumption that the amplifiers operate as transconductance amplifiers is to be valid, the external load impedance Z_L must be much smaller than the internal output parallel impedance $Z_O = 1/g_{mo}$ (see figure 2.7). In the transconductance-C design Z_L is purely capacitive, and hence the effective load impedance will be very large for low frequencies, but insignificant for frequencies much larger than g_{mo}/C, where C is the load capacitance.

There are two major differences between equation 2.1 and the corresponding equation in [41]. The first is that the new expression takes damping in the pass band into consideration, which is useful for computation of the signal strength throughout the cascade. The second difference is that the cutoff frequency and Q-factor expressions include information about the effect of C_1 and C_2 in combination with the output conductance g_{mo}. This information is useful for selection of C_1 and C_2 values to make ω_0 and Q approach what is expected from the simplified equations 2.9 and 2.10.

Nonlinearities

The selected OTA amplifiers (see section 2.5) contribute with two particularly interesting nonlinear effects. The first is nonlinearity in the OTA transfer function. The g_m typically

exhibit a hyperbolic tangent function instead of a strictly linear function. Second the OTA use a back-gate input using the wells of the differential input pair to reduce the gain of the OTA. This increases the linearity of the g_m characteristic by reduction of the amplifier gain. Unfortunately the wells create a depletion region in the PN junction between the substrate and well with thickness dependent on the junction potential, thus a significant non-linear parasitic capacitance is created. To reduce the effect of this well-junction capacitance, main load capacitances of the SOS section is created by poly capacitances. But to maintain a modest current and space consumption, fairly small poly capacitances are used. Thus some of the nonlinear effect will be evident for the entire frequency range.

The g_m nonlinearity is mainly effective around and beyond the best frequency point. When the input differential voltage is small, the g_m is in the most linear region. But at the resonant frequency and beyond, the differential input voltages are out of phase, creating a large input to the OTA pushing g_m towards the nonlinear regions. Thus this effect should only be evident around BF, and looking down the cascade the effect should saturate at a level decided by the number of overlapping resonant peaks creating the BF of the observed output. Reduction of the g_m nonlinearity is treated later in section 2.5.

Dimensioning

Inclusion of the damping factor in the general equation for frequency response of a second order section from [33], give:

$$|H(\omega)| = \frac{\delta}{\sqrt{[1-(\omega/\omega_o)^2]^2 + 4k^2(\omega/\omega_o)^2}} \tag{2.11}$$

where δ is the damping factor and ω_o is the undamped resonance angular frequency according to formula 2.3 and 2.4. The quality factor is computed according to formula 2.5 and included in the formula above with

$$k \equiv \frac{1}{2Q}$$

Considering the effect of the damping factor on the filter response, equation 2.3 show that the damping is given by the output conductance and amplifier transconductance. Simulations show that the output conductance is typically $4.5\text{n}\mho$, and g_m might be varied between $1\text{p}\mho$ and $10\mu\mho$. From equation 2.3 it can be shown theoretically that the damping factor will vary from 1 to $494n$, or in other words from 0dB to -146dB. Although the latter number is probably exaggerated since not all parasitic effects are considered in this formula, it indicates that the complete signal will be lost. To counteract this effect, sufficiently large g_m must be provided by the amplifiers, and the minimum cutoff frequency must be decided with sufficiently large capacitors.

From equation 2.11 it might be argued that the resonant peak of the filter can be used to counteract the low frequency damping. This is true for frequencies near the resonant peak, which might still be amplified despite the damping factor. But for a cochlear cascade the first filter sections have a very high cutoff frequency, and it is easy to show that an unreasonably large resonance is necessary to avoid damping of the low frequency components. This problem is even worse when the number of stages in

the cascade increases causing a slow decrease in the cutoff frequency compared to the number of stages down the cascade. But still the resonance reduces the overall damping, making it a bit easier to achieve a satisfactory SNR.

The back-gate OTA input transistors exhibit substantial parasitic junction depletion bulk-source and bulk-drain capacitances, creating an internal feedback on the input capacitances. To make sure the parasitic capacitance is well below the load capacitance, relatively small transistors should be used. If the parasites start to dominate, the parasitic feedback interferes at the cutoff frequency which in turn increases the filter resonance.

The small transistors and the linearization techniques generate a relatively low g_m. Thus to be able to operate the filter sections at high frequencies, large current scaling is implemented in the current mirrors to increase g_m. This scaling permits high maximum cutoff frequency, which must be 100kHz when $\omega_1 = \omega_2$ if Q should be adjustable in the interval $[0.2, 5]$.

Damping is reduced both by the cascode transistors at the output (which reduce g_{mo}), and sufficient multiplication in the current mirrors to increase the g_m while the amplifier is permitted to operate in sub-threshold over the entire frequency range. By multiplication in the current mirrors, the linearity of the core can be increased by reduced transconductance in the core, while a satisfactory gain is maintained for the complete OTA.

As mentioned earlier, to reduce the nonlinear effects of the parasitic capacitances, massive poly capacitances must be added. Unfortunately this consume much space, typically $2 * 100 * 100\mu m^2$ for each SOS section to reduce the effect of the parasitic capacitances to an ignorable level. From general size considerations, it was decided that poly1-poly2 capacitances of $1.4pF$ $(40\mu \times 40\mu)$ and $1.0pF$ $(40\mu \times 30\mu)$ should be added to the parasitic capacitances. This capacitance is in the order of one magnitude larger than the parasitic capacitance, and the different values are selected to somewhat balance out the difference in the parasitic load, and equal the total values of C_1 and C_2. The main parasitic load is expected from the well substrate junction of the following second order stage, but the value is difficult to estimate because it is not given in the process parameters. Other parasites are at least two orders of magnitude smaller than the capacitive value of the poly capacitors, and are ignored.

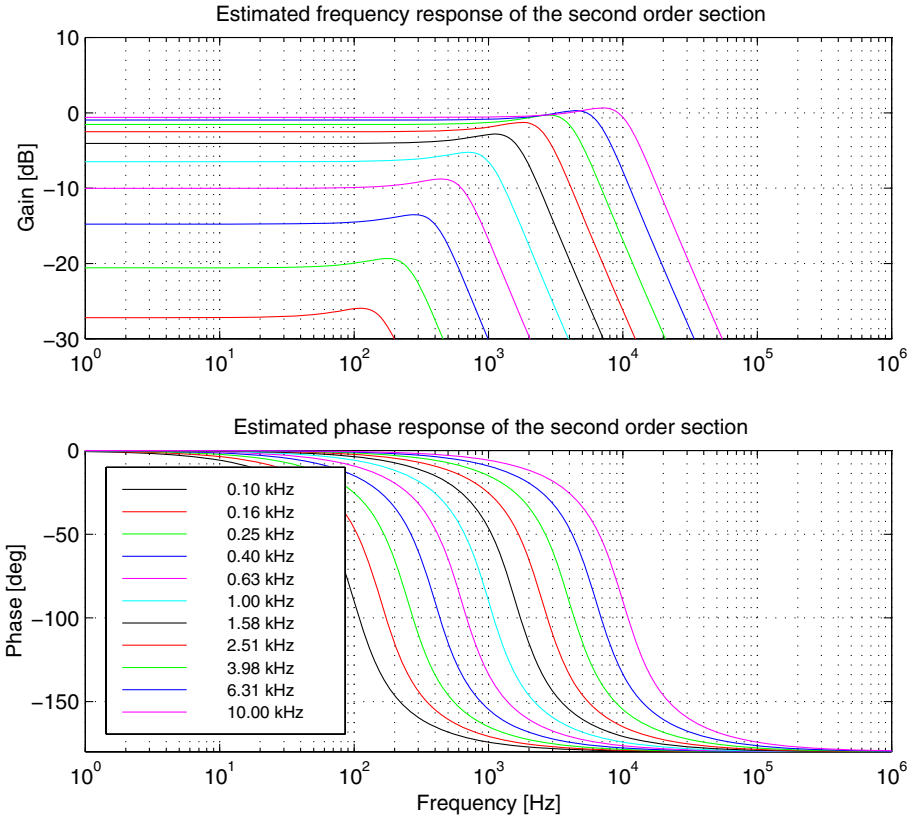


Figure 2.8: Estimated frequency and phase response of the second order section with $Q = 1$ and varying f_o using MATLAB.

In figure 2.8 the result from a MATLAB simulation using formula 2.3, 2.4, 2.5 and 2.11 is shown. The necessary g_m for a given ω_o and Q is computed from a new set of equations found from the simplified equations for ω_o and Q. The two equations are based on the assumption that $g_{mo} << \{g_{m1}, g_{m2}\}$ and is found as:

$$g_{m1} \approx \omega_0 Q C_1 \tag{2.12}$$

$$g_{m2} \approx \frac{\omega_0 C_2}{Q} \tag{2.13}$$

The deduction of these equations can be found in A.4 in appendix A. In the simulations $Q = 1$, $C_1 = C_2 = 1.5pF$ and $g_{mo} = 1n\mho$.

The substantial damping for the lower frequencies is in correspondence with earlier assumptions, and is large for the intended application. But as shown later, higher order effects counteract some of the damping, and reduce it to an acceptable level. Note the compression of the actual ω_o due to the increasing influence of g_{mo}. The phase response is as expected of a second order low-pass filter.

2.5 OTA implementation

The second order filter sections are based on the weak inversion amplifiers from [41, 42]. By using the transistors in weak inversion, very low current consumption is achieved. This is both beneficial for the total power consumption budget and area consumption because small capacitive load is necessary to achieve the desired cutoff frequencies. With the exponential g_m dependence of the bias voltage acquired with the sub-threshold operation, the same filter section might be used over the entire cutoff-frequency range of the cascade filter.

A general drawback with sub-threshold amplifiers is that there exist few linearization techniques. When using above threshold amplifiers general square law relations can be used, but in weak inversion the transistors operate in the exponential domain. Thus instead linearization methods from BJT OTA design can be adapted. Emitter degeneration, described in [9], is the standard solution which increase the linear range by reduction in the transconductance. Thus the hyperbolic tangent transfer function of the OTA is flattened increasing the length of the close to linear center characteristic. Fortunately low transconductance is beneficial for the second order filters because low cutoff frequencies are made achievable with small capacitances. In the design three different linearization techniques are used, where two use transconductance reduction.

As discussed later in section 2.6, good linearity is important to keep the output of the second order filter sections from drifting and to reduce distortion of the signal.

2.5.1 OTA architecture

One current mirror and the current source of the amplifier have been improved, both to make sure the amplifier is operating correctly, and to improve its general performance. Generally when dimensioning the transistors, the original discussion from [42] has been followed for the parts not discussed below. The complete schematic is shown in figure 2.9.

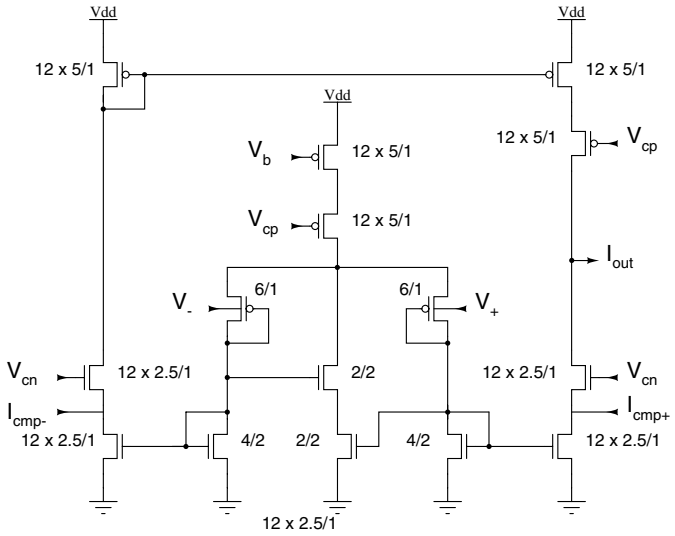


Figure 2.9: Wide range OTA used in the basic filter section of the filter cascade. The I_{cmp+} and I_{cmp-} inputs are the current inputs from the comparator. Note the back-gate input of the diff-pair and the bump linearization transistors between the source diff-pair and ground.

Linearization is partly achieved with the two PMOS differential input transistors shown in the center of the figure, which use gate degeneration and bulk input. By utilization of the transistor current dependency of the bulk voltage, the transistor transconductance is reduced. The current through the transistor is dependent both by the gate and well voltage according to

$$i_{DS} = I_0 exp(-\frac{\kappa v_{GS}}{U_T}) exp(-\frac{(1-\kappa) v_{WS}}{U_T}) \tag{2.14}$$

where v_{GS} and v_{WS} are the gate-source and well-source voltages, U_T is the thermal voltage, I_0 is the sub-threshold current factor and κ is the sub-threshold exponential coefficient [42]. Usually $\kappa > 0.5$, making the transconductance of the well smaller than the gate. A common problem with the back gate input is large parasitic capacitance, but for our application this capacitance will added to the integrator capacitance. It should be noted however that this capacitance is relatively nonlinear and will introduce some harmonic components. To reduce the impact of the parasite, additional capacitance has been added to the filter, as described earlier.

The second method for transconductance reduction is gate degeneration by connection of the transistor gate to its drain. When the drain-source current increase, the drain-source and hence gate-source voltage diminishes. This creates a further transconductance reduction.

Finally the two center NMOS transistors linearize the response of the differential pair using a technique called bump linearization [20, 41, 42]. These transistors create a bump around the origin in respect to the differential input voltage. This bump steals

current from the differential pair, thus flattening the usual tanh characteristic around the origin if the transistors are properly dimensioned.

As discussed in section 2.4, the output transconductance, g_{mo}, is very important to maintain a sufficient SNR ratio through the complete filter cascade. To reduce this, the output currents are mirrored with cascode current mirrors. These cascode transistors greatly reduce the early effect of the output stage, and in effect the g_{mo} is reduced to a tolerable level.

Two structural changes have been made to the original WLR-amplifier design. First the PMOS cascode transistor on the input of the output PMOS current mirror is removed. With this transistor the gate-source voltage of the diode coupled input transistor is reduced causing the output transistor of the current mirror to supply too little current. With this transistor in place the current mirror will not work with a reasonable cascode voltage. If $V_{cp} = 0V$, the mirror will work, but all PMOS cascode transistors are turned fully on and will have very little effect. To further improve the linearity of the amplifier, a cascode transistor has been added to the current source transistor. This reduces the general voltage dependency of the current source and thus makes the amplifier current less dependent of the common mode voltage of the differential pair. The drawback is less tail current swing on the differential pair reducing the headroom before the OTA starts clipping. But the main limitation of the OTA is the linear range, thus even though the total dynamic range of the amplifier is reduced, the usable dynamic range with satisfactory harmonic distortion is increased.

2.5.2 Dimensioning

From the formulas in section 2.4 the necessary estimated transconductance range is $<$ 370n℧, 470p℧ $>$ to achieve a frequency range of $<$ 100Hz, 20kHz $>$ and simultaneously a Q-range of $< 0.5, 2 >$. The necessary output transconductance g_{mo} is more difficult to estimate because it is dependent of the amplifier gain which is dependent of the g_m of the transconductors which varies through the cascade. Using worst case numbers, $g_{mo} <$ 940f℧ for less than 0.05dB/filter, or a total damping of 5dB at the last filter cascade stage. If instead damping is extracted from hspice simulations for different cutoff frequencies, a $g_{mo} <$ 1n℧ is sufficient when the lowered damping for high cutoff frequencies are considered.

To permit comparison with current implant processors, the dynamic range should preferably equal the range of the reference device. Further the dynamic range should preferably be large enough to not significantly affect the perceived sound quality. Currently the dynamic range is approximately 60dB of commercial available implants, thus the dynamic range of the amplifiers should be at least this large.

The transistors in the current mirrors and the current source have been generously dimensioned, and there is room for space reductions to reduce parasitic capacitance and space consumption. Transistors in the current-mirrors are made large to improve matching. The core current is scaled by a factor 15 in the mirrors to match the g_m requirements of the second order filter section.

2.6 Offset correction circuit

An inherent problem with second order sections using the architecture described in section 2.4, is a DC-offset drift on the output node. The main source of this problem is probably nonlinearities in the amplifiers (g_{m1} and g_{m2} in figure 2.10). The input differential voltages of the amplifiers increase with increasing lag in the voltage across the output capacitors when the filter is operated above its cutoff frequency. This difference causes the amplifiers to supply a nonlinear current instead of the ideally linearly proportional output current to the input voltage difference. The current will typically be slightly larger on the positive half period of the input signal and similarly reduced during the negative half period of the input signal. Over time this creates a positive excess charge on the output capacitances and the voltage drift away from the ideal DC-value.

To reduce this problem, the offset correction circuit from [41] is implemented. This circuit extracts the DC value of the SOS section (see figure 2.4) and compares it with a reference voltage. In response to a difference the comparator injects current in the outer current mirror arms of the OTA (see figure 2.9) to correct the DC value.

With the original first order filter, the signal is not enough filtered before it is feed back with the correction current. Unfortunately this creates unwanted distortion of the signal, particularly for low frequency signals. Because low-frequency signals are not very well damped and are traveling through most of the filter cascade, this affects a large number of OCR circuits. A reduction of filter cutoff frequency might be attempted to resolve the problem, but alas the filter section area consumption will increase radically since the capacitive load must be increased to maintain a sufficiently large OTA operation current. But perhaps worse is that the response time of the OCR increases as the cutoff frequency is lowered.

To solve this problem without compromising the response of the OCR circuit, a second order filter section is used as described in article [26, 27] included in part II. Thus the stop-band damping is increased even though the cutoff frequency is held constant to maintain a sufficiently fast response to changes in the DC offset level.

2.6.1 Deciding filter order

Several considerations apply when deciding the order of the correction filter. High order creates a sharp transition from pass-band to stop-band, but increases the offset of the correction filter. Furthermore, a sharp transition can be exploited to increase the cutoff frequency still keeping sufficient signal rejection, hence permitting smaller capacitors and reducing the total area consumption.

If the cutoff frequency of the OCR filter is set to 10Hz, fast adaptation to varying offsets and acceptable capacitor sizes are achieved. The lowest cascade filter signal-frequency set to 100Hz. With these numbers a 2nd order correction filter offers 40dB damping of the lowest signal frequency, which is a little less than the expected dynamic-range. But it is probably sufficient, bearing in mind that the damping in the more information-carrying part of the sound picture around 1kHz is 80dB, which is beyond the expected dynamic range of the cascade. In comparison a first order filter will only reduce the signal frequencies half as much.

2.6.2 Design

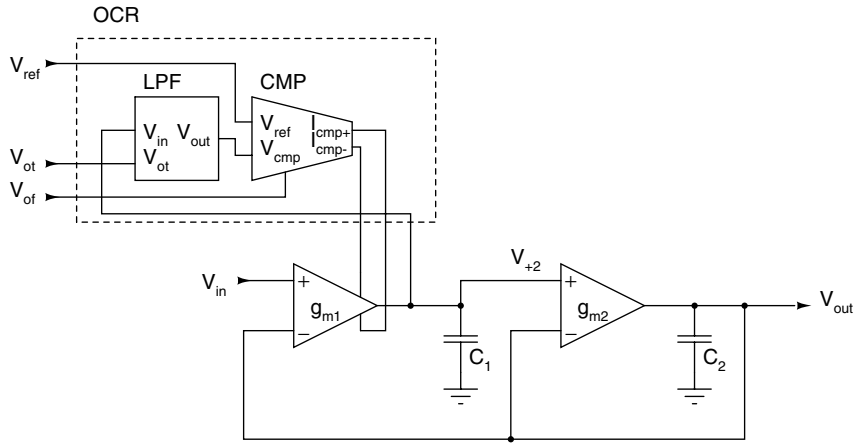


Figure 2.10: Block-diagram showing one second order section with offset correction.

In figure 2.10 a block diagram show the connection of the offset adaptation circuit to a second order section[1]. In the diagram the LPF section extracts the DC voltage over C_1, which is then compared with a reference voltage V_{ref}. If the extracted voltage differs from V_{ref}, current is feed into the SOS section to adjust the DC-level over C_1. Note that the input reference circuit described in [26, 27] has not been implemented on the final chip to reduce test chip complexity and focus on the central parts of the design.

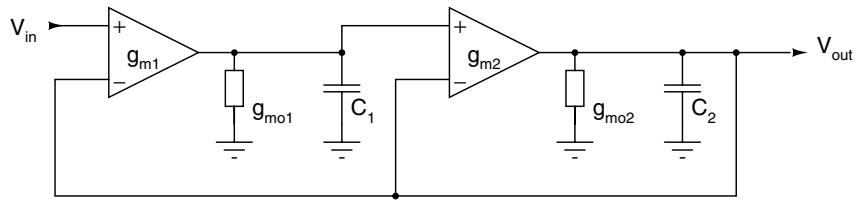


Figure 2.11: General architecture used for the second-order low-pass section. The conductance in parallel with the output capacitor is the output conductance of the amplifiers. Note the similarities to the architecture of the filter section described in section 2.4.

The internals of the low-pass filter is shown in figure 2.11. The general architecture is identical to the SOS filter architecture described in section 2.4, but the OTA´s have been replaced with the simple OTA in figure 2.12 to reduce area consumption. Although the linearity is poor compared with the WLR amplifiers, the resulting offset of the OCR filter is much smaller than the cumulative offset of the cascade.

The comparator, shown in figure 2.13, compares the extracted DC-voltage (V_{out}) with the reference voltage (V_{ref}). If the voltages differs, current is injected into the

[1]Note that the design blocks have been renamed with respect to the original design in [41].

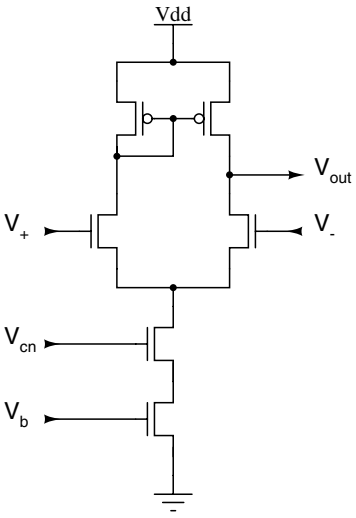


Figure 2.12: Operational transconductance amplifier used in the low-pass filter section of the offset correction filter.

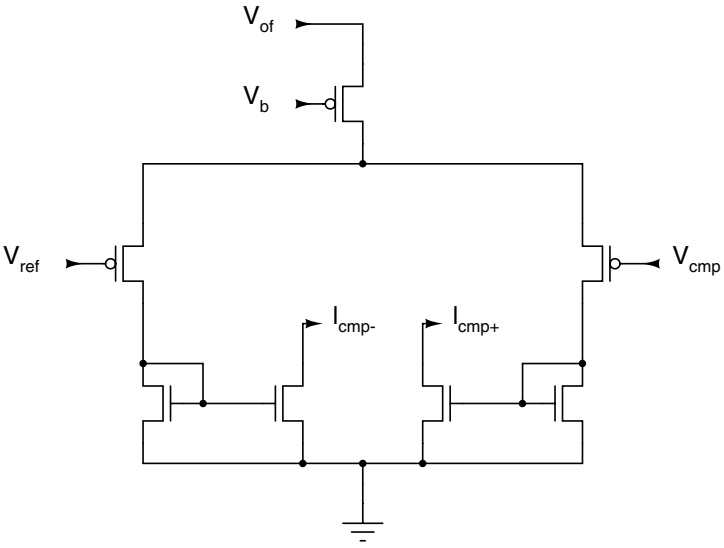


Figure 2.13: Comparator used in the offset correction circuit.

outer current mirrors of g_{m1}, marked I_{cmp+} and I_{cmp-} in figure 2.9 in section 2.5. The V_b bias is connected to the input bias of g_{m1} and is used to automatically scale the size of the injected currents when the bias of the OTA is adjusted. The V_{of} bias is constant during normal operation of the circuit and adjust the absolute current injected in g_{m1}.

2.7 Simulation results

Numerous simulations have been performed to validate all parts of the design, but even to optimize design parameters. In the following subsections a selection of the most important simulations are described.

2.7.1 Second order section

An AC simulation in Spectre has been run on the design to verify that it is functioning in the desired frequency range. With V_{b1} and V_{b2} equal and varying from $[4.6V, 4.3V]$ the undamped resonant frequency varies in the interval $[15Hz, 140kHz]$. Because the biases are equal, Q should be constant and equal one for all the settings if the damping factor is ignorable and the capacitances are perfectly balanced.

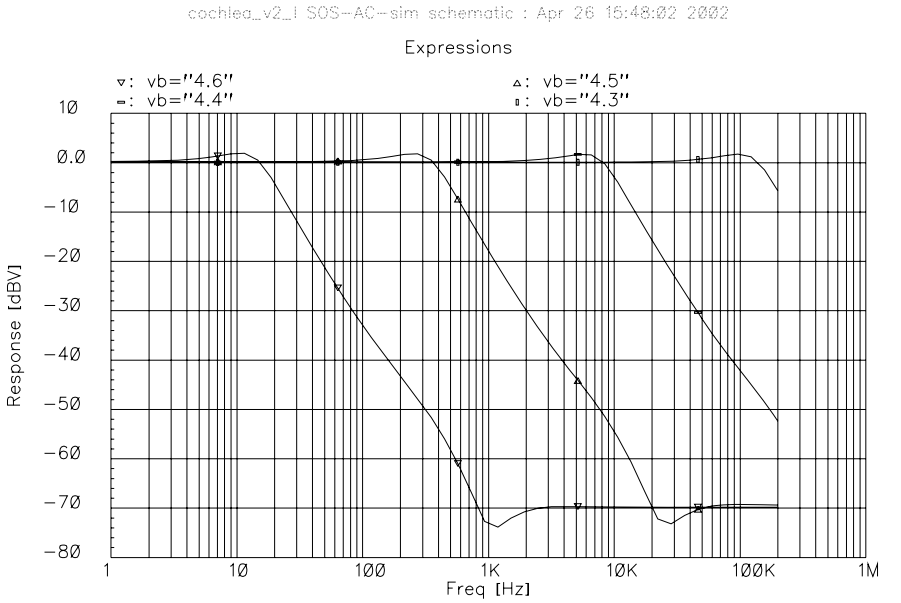


Figure 2.14: Simulated frequency response of the second order section (Spectre simulation on the schematics).

The cascode bias voltages are $V_{cp} = 3V$ and $V_{cn} = 2V$. Input common mode voltage is $3V$, and AC amplitude is $500mV$. Compared to the MATLAB simulation in section 2.4.1, the simulation in figure 2.14 show good frequency response and ignorable

damping. This is probably either due to a pessimistic g_{mo} estimation used in the previous MATLAB simulation, or signal leaking through the parasitic connection from the bulk input at the WLR amplifiers to the internal drains at the differential pair. From the Spectre simulation it seems that the third order effects are less dominant because the roll off is approximately $40dB/dec$ as expected of a second order section.

One final difference is even evident as the maximum value of the transfer curve is two to three decibels larger than the expected 1.25dB when $Q = 1$, which supports the theory of internal parasitic coupling in the WLR-amplifiers.

2.7.2 Wide linear range amplifier

The complete cochlea design is based upon the WLR amplifier, thus its functionality has been thoroughly verified with simulations before the chip was produced. Since there are no direct measurements on the amplifier, some of the simulation results are presented her to qualify the amplifier.

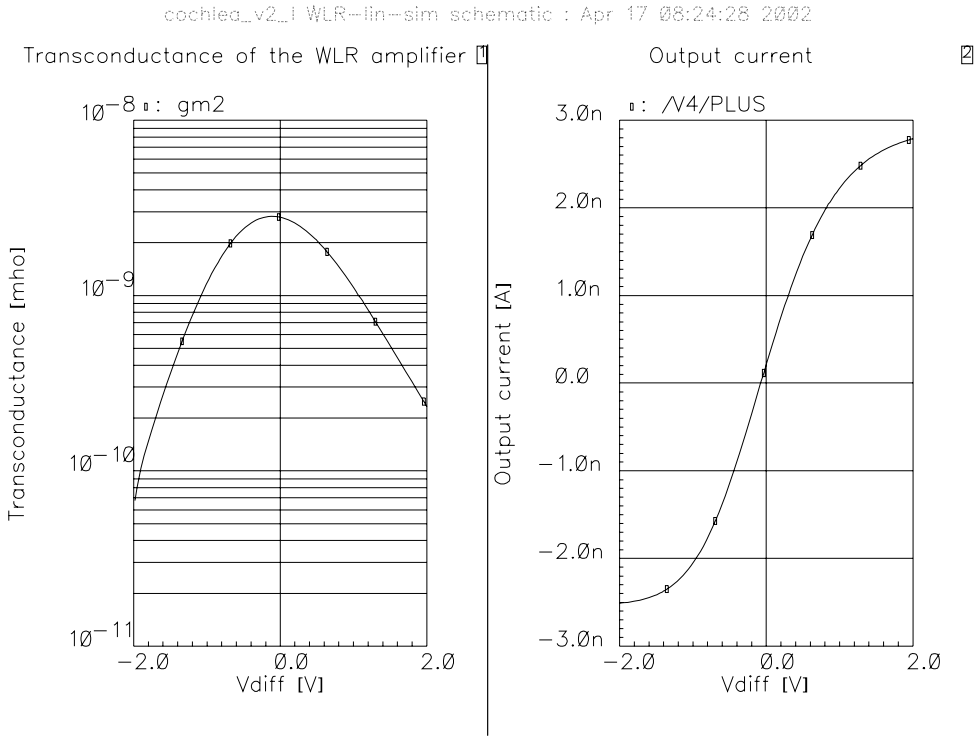


Figure 2.15: Simulation showing the computed transconductance of the WLR amplifier and its output current.

Figure 2.15 show the results of an amplifier simulation. The left plot show the variation of the amplifiers transconductance, g_m, computed as the derivative of the output current with respect to the input differential voltage. Figure 2.16 show the

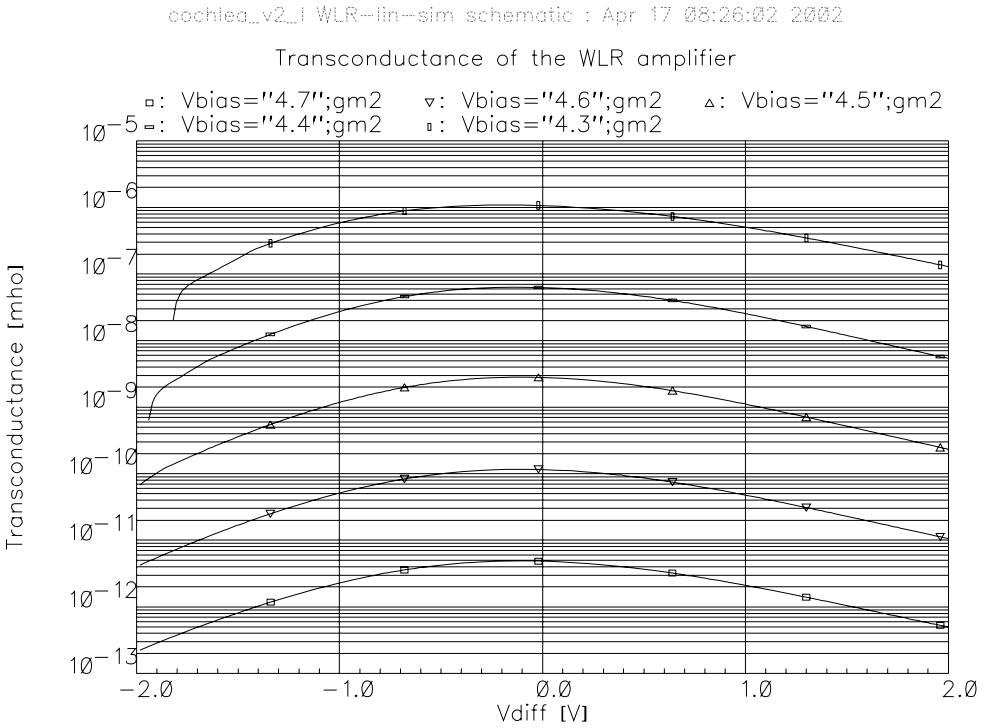


Figure 2.16: Simulation showing the computed transconductance of the WLR amplifier for different bias voltages.

Parameter	**Value**
Output voltage	3.0V
Input common mode voltage	3.0V
V_{cp}	3.0V
V_{cn}	2.0V
V_{bias}, single run	4.5V
V_{bias}, multi run	4.3V, 4.4V, 4.5V, 4.6V, 4.7V

Table 2.1: Parameters used in the simulation of figure 2.15 and figure 2.16.

transconductance when the bias voltage is varied. Table 2.1 show the parameters used for the simulations.

From the first plot a linear range with maximum 1% transconductance deviation is found to be approximately 0.20 volts. From the second graph, it is evident that a current source bias in the range $[4.3V, 4.7V]$ give $g_m = [1.1\mu\mho, 5.4p\mho]$. All results are satisfactory for the intended application.

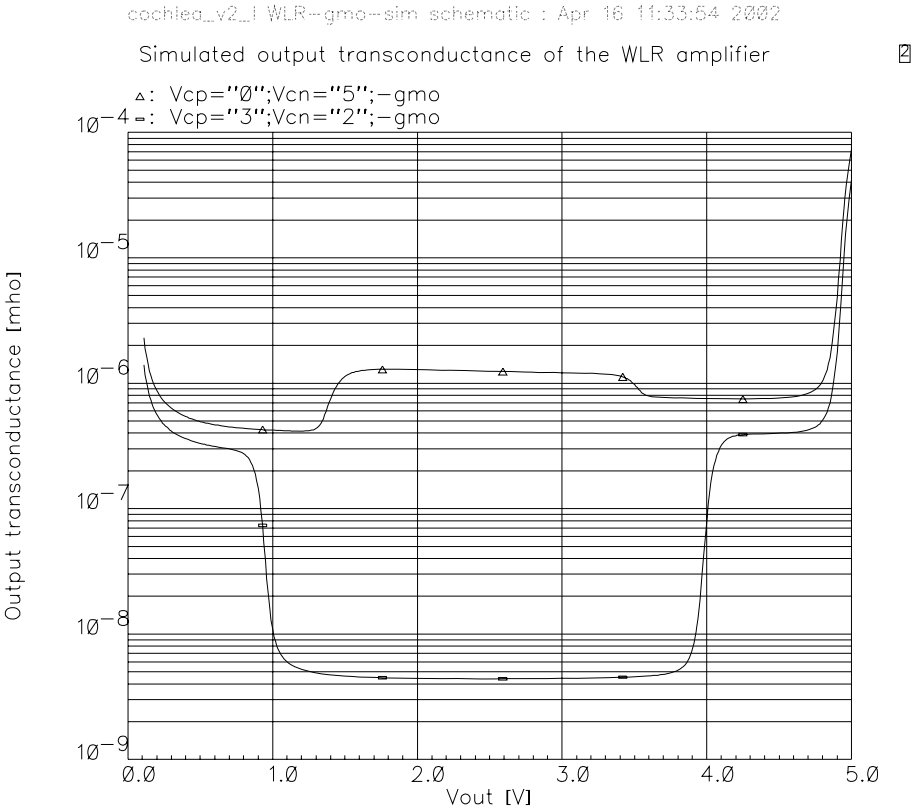


Figure 2.17: Simulation showing the computed output transconductance of the WLR amplifier with and without active cascode transistors.

In figure 2.17 the results of a g_{mo} simulation is shown. One run is with the cascode transistors in function, and one is with the effect of the cascode transistors minimized by setting the bias-voltages to $V_{cp} = 0V$ and $V_{cn} = 5V$. Table 2.2 show the parameters used in the output transconductance simulation. The simulations show both a general reduction of g_{mo} by almost three decades, a slightly more constant g_{mo}, and larger linear range when the cascode transistors are enabled.

Parameter	Value
Input voltage	0V
Input common mode voltage	3.5V
V_{cp}	3.0V
V_{cn}	2.0V
V_{bias}	4.3V

Table 2.2: Parameters used in the simulation of figure 2.17.

2.7.3 Offset correction section

Since the low pass filter is well documented in section 2.4, only a frequency response simulation has been included to verify that the filter can be adjusted within a reasonable frequency range.

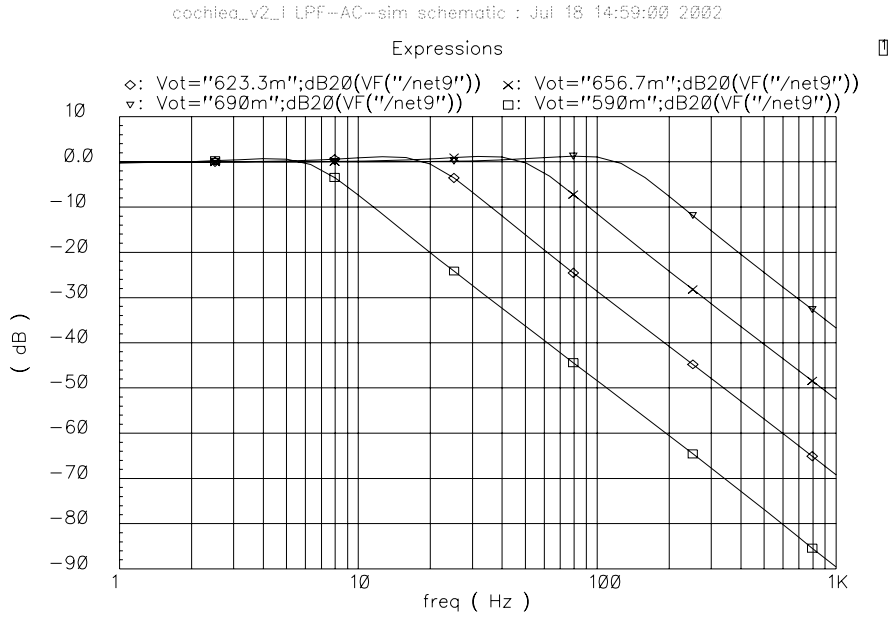


Figure 2.18: Simulated frequency response of the second-order section for offset correction.

Figure 2.18 show the results from a parametric Spectre simulation on the filter section where the bias voltage has been adjusted between the runs. The results clearly show very good headroom for frequency adjustment and the expected 40dB/dec decrease in the stop band for a second order section.

2.8 Measurements

A number of measurements have been carried out on the cochlea and its sub circuits to establish whether the cochlea is working properly for the desired operating ranges, and to verify the analysis of the second order section explained in section 2.4. But the WLR amplifier is considered to be well tested by measurements on the filter section, and to reduce chip complexity no separate test structure was added to the test chip for this purpose. Thus by the results found for the SOS blocks the WLR is indirectly verified.

2.8.1 Measurement setup

The measurements are limited to five buffered outputs which are directly connected to pads. There are however another option as all the outputs are connected to a separate neuron which all can be separately configured via the switch matrix to connect to an output pad, but as explained later in chapter 3, the voltage-to-current converters are extremely sensitive to processing parameters, causing the spike output of the neurons to be greatly affected. In addition the neurons are affecting the measured response by inclusion of quantization noise. Thus to make the measurements reasonable accurate, only the five accessible direct outputs of the buffers have been used. It was also necessary to use a measurement device with very low loading of the outputs due to poor driving capabilities of the buffers.

To achieve a reasonable output load, a Lecroy 9450 oscilloscope was used for data collection, transferring the data to MATLAB via a GPIB interface. Gain measurements have been performed by reading the RMS-voltage of the cochlea input and output signal, and then computing the corresponding dB value in MATLAB. Phase, THD and SNR have been computed with the standard *fft* function in MATLAB using the measured waveforms from the oscilloscope. The phase was then unwrapped with a script using the *unwrap* function in Matlab, expanding negative phase jumps larger than 2π. THD was computed as the ratio of the sum of the powers of all harmonic frequencies[2] of the output signal and the power of the fundamental frequency. SNR is computed as the minimum difference between the output signal and all noise components at least 50% away from the signal frequency to avoid interference from the main and nearest side lobes of the signal. The phase is computed using the *angle* function in MATLAB. Note that the phase collapses for very steep phase decline, which is due to limitations in the unwrap function.

Filter time constants τ and quality factor Q was set by the ratio of the bias currents of the SOS blocks using a Keithley 213 controlled from MATLAB using a script computing the necessary bias voltages for given Q and τ for the first and last filter section. The script was based on the simplified formulas 2.10 and 2.9 from section 2.4.

In all the measurements the input signal was a pure sine wave with 500mV amplitude and 2.5V offset, unless otherwise stated. It was generated by an Agilent 33250A arbitrary function generator where the input frequency was controlled from a MATLAB script via the GPIB interface.

[2]The number of computed harmonic components is limited to approximately five by the range of the FFT-transform and the placement of the fundamental frequency.

2.8.2 Cochlear cascade

For testing of the filter cascade response without cascode transistors, the circuit has been measured with maximum cascode bias voltages reducing the cascade effect to an ignorable level.

Generally the lower limit of the signal response in the plots results from the limited range of the oscilloscope, and not the cochlear cascade. Using similar settings as in the plots, the maximum measurable attenuation in the frequency response plots has been confirmed to be approximately -30dB. Note that different settings were used in the Q-factor measurements.

Frequency response

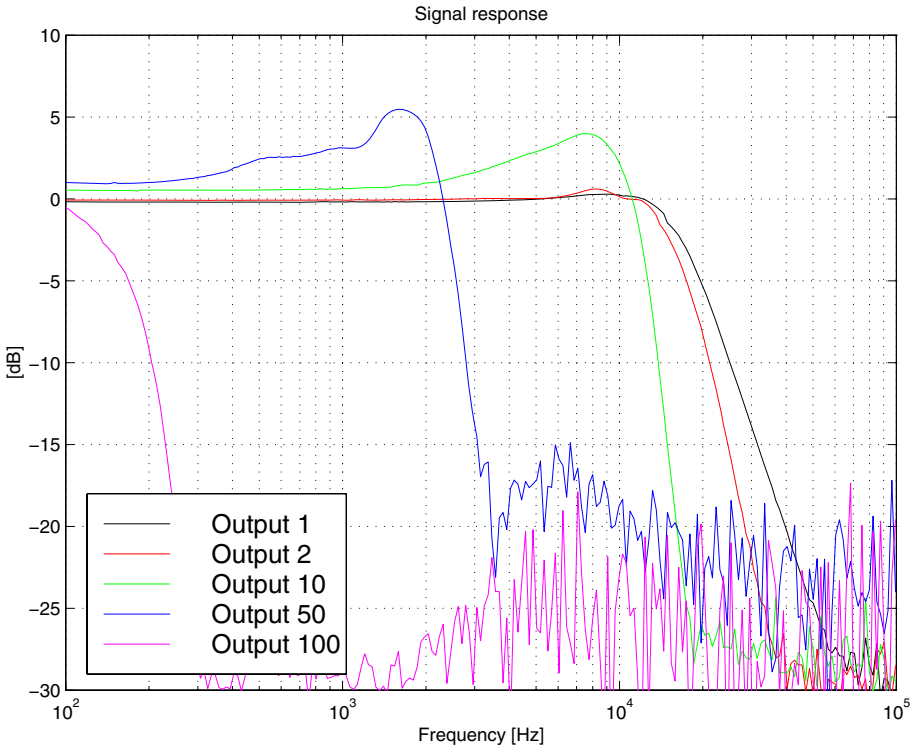


Figure 2.19: Measured frequency response on cochlea output taps.

Figure 2.19 show the gain-frequency response for typical settings, while figure 2.20 show a corresponding plot with the effect of the cascode transistors minimized.

In the plots the undamped resonant frequencies were set to 15kHz at the base and 200Hz at the apex using the formulas described in section 2.8.3. According to the definition the expected undamped resonant frequency (f_0) of second order filters [33], f_0 is the frequency of the filter response at 0dB crossing when Q=1 and $\delta = 1$. If Q

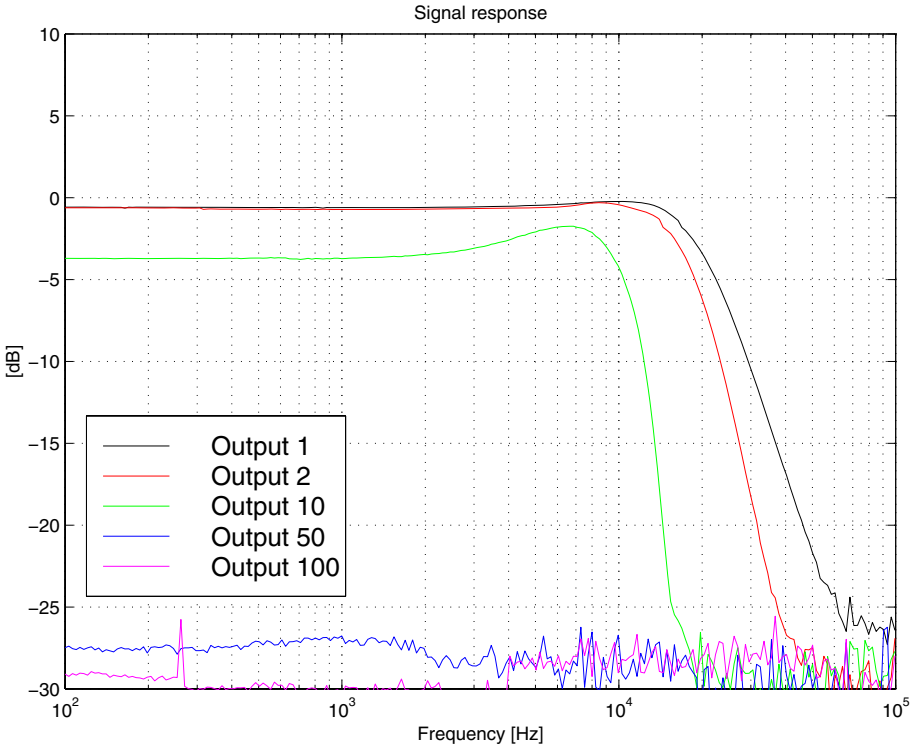


Figure 2.20: Measured signal response on cochlea output taps with minimized cascode transistor effect.

is reduced, the 0dB crossing moves to lower frequencies even if f_0 is constant. But in the cascaded sections the output of the observed filter is a multiply of the response of the preceding sections with higher undamped frequencies. Thus the effective undamped frequency is increased because the signal is amplified by the preceding sections. This result corresponds well for all outputs except output 100 where the damping factors dominate such that the effective undamped frequency is reduced.

According to the earlier discussion, the pseudo resonance should cause an increase in the signal gain, particularly for the lower frequency taps, as the response of the frequency peaks accumulate. When comparing output one in figure 2.19, which is from a single filter section, with the other outputs it is clear that the theory is correct. The resonant peaks are increasing sharply up to output 50 from almost 0dB to approximately 6dB. But at output 100 the peak is missing, most likely because the output transconductance is large enough to cause significant damping for the low frequency components. Note that Q is constant for all SOS blocks through the cascade. See even equation 2.3 in section 2.4.

From the results with reduced cascade effect in figure 2.20, a strong cumulative damping of the signal in the pass band is clearly evident, which is as expected from equation 2.3. According to equation 2.7 the undamped resonant frequency should be

slightly lowered due to increased g_{mo}, and the Q-factor should decrease slightly according to equation 2.5. This behavior is evident as the resonant peaks are slightly attenuated compared with the original plot. Although it is difficult to tell if the resonance frequency is lowered, a reduction from approximately 7.5kHz to 6.8kHz can bee seen on output ten where the effect accumulates and the signal is still strong enough to detect the peak.

Phase response

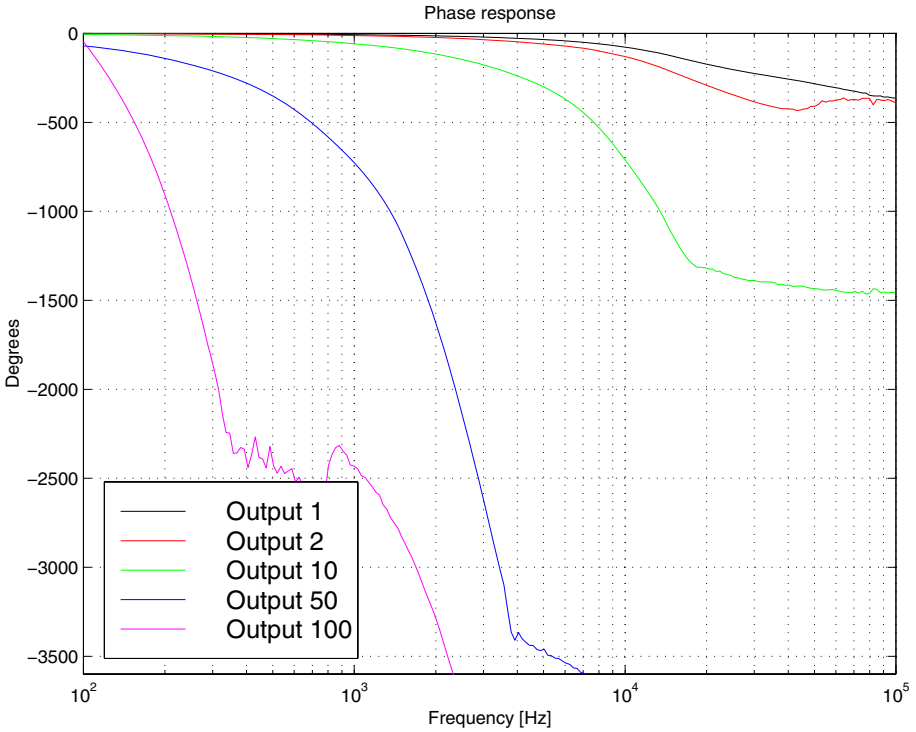


Figure 2.21: Measured phase response on cochlea output taps.

The earlier mentioned phase delay due to signal propagation through the cascade can be seen in figure 2.21. Particularly output 50 have a very large delay of about 1300 degrees, or 3-4 periods at the resonant peak. On the high frequency outputs fewer filter sections contribute and the delay is clearly much smaller. For lower frequencies the total delay should approach a constant value since the first filter sections of the cascade have a much higher cutoff frequency than the signal, and thus contributes little to the phase lag. Even though output 100 is unreliable, the phase response is reasonable. The distortion evident at the end of the phase response curves is created by limitations in the phase unwrapping algorithm and are caused either by phase jumps larger than 2π or by phase increment.

Signal to noise ratio

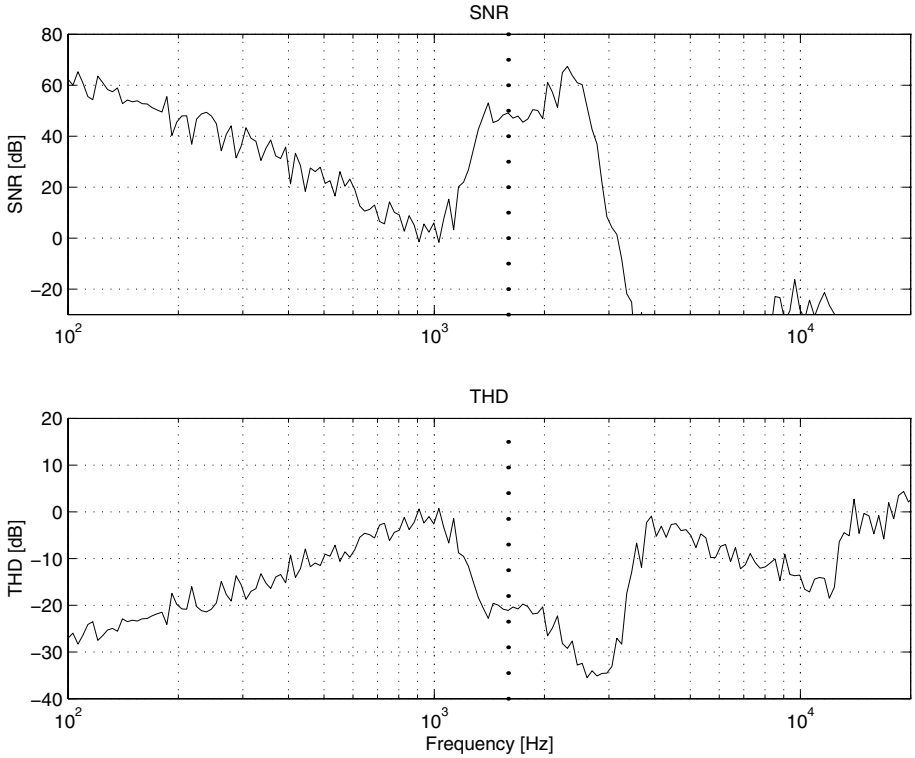


Figure 2.22: Measured SNR and THD on cochlea output 50 with the cascode transistors enabled and 250mV signal amplitude. The dotted vertical line indicates the best frequency for the current output.

Figure 2.22 show SNR and THD of output 50 relative to the input frequency. The dotted vertical line indicates the center of the resonant peak of this output extrapolated from the maximum value of the response given in plot 2.19, or the best frequency (BF) of this output.

The negative SNR slope up to approximately BF/2 is caused by harmonic noise. The dominant harmonic noise appears as damped copies of the signal frequency where $f(n)_h = 2^{n-1} * f_s$, $f(n)_h$ is the frequency of harmonic component n and f_s is the signal frequency. The strength of the harmonic components is inversely proportional with its number (n). Thus when the signal frequency is below the resonant peak, the harmonic components are amplified while the signal component is constant. And when the signal frequency is increased, the lower order and hence stronger harmonic components are amplified, further reducing the SNR. But beyond BF/2 the harmonic noise is damped by the sharp filter response while the signal is amplified by the resonant peak, dramatically improving the SNR ratio. The increasing THD in the lower plot supports this analysis.

Figure 2.23 show similar results with lower input amplitude, reducing the harmonic

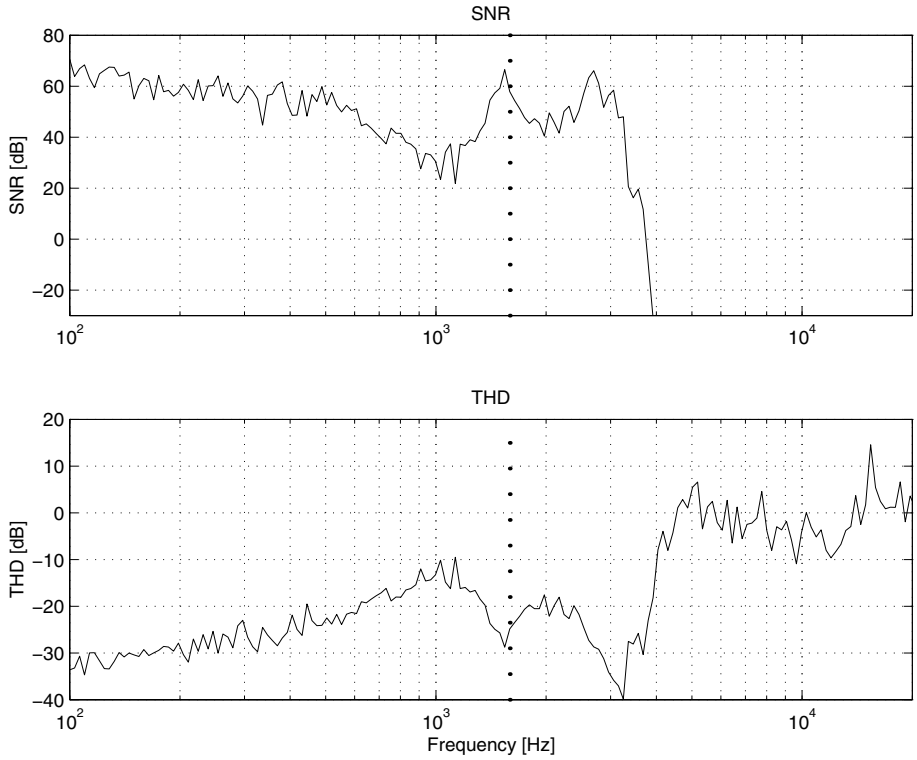


Figure 2.23: Measured SNR on cochlea output 50 with the cascode transistors enabled and 50mV signal amplitude. The dotted vertical line indicates the best frequency for the current output.

distortion. As expected the response is flatter and the worst SNR in the passband is increased from approximately 0dB to more than 20dB. The SNR at the BF is increased from about 50dB to 60dB, partly explained by less harmonic distortion and partly by more headroom permitting a larger increase in the output amplitude before the OTAs saturate.

A second SNR peak at approximately 2.7kHz is evident as the last harmonic peak is suppressed, thus the most dominant noise component is practically removed. This peak is even stronger in figure 2.22, further verifying the assumption that harmonic noise affects the measurements.

Table 2.3 show SNR for the best frequency (BF) at all measured outputs. The SNR of output 50 exceeds the results of output 30 in [41] with 20dB.

At output 50 the SNR is greatly improved when the cascode transistors are enabled. But for output 1 to 10 the distortion gets slightly worse. This can be explained by reduced linear range creating harmonic noise and reducing the signal amplitude. For the high frequency outputs this effect is dominating over the increased pass band damping when the cascode transistors are missing.

Output	Full cascode effect	Minimized cascode effect
1	65.5 dB	74.3 dB
2	55.6 dB	61.3 dB
10	39.5 dB	44.7 dB
50	51.5 dB	23.8 dB
100	13.4 dB	-26.3 dB

Table 2.3: Measured SNR at the best frequency (BF) of the measured outputs.

Q-factor

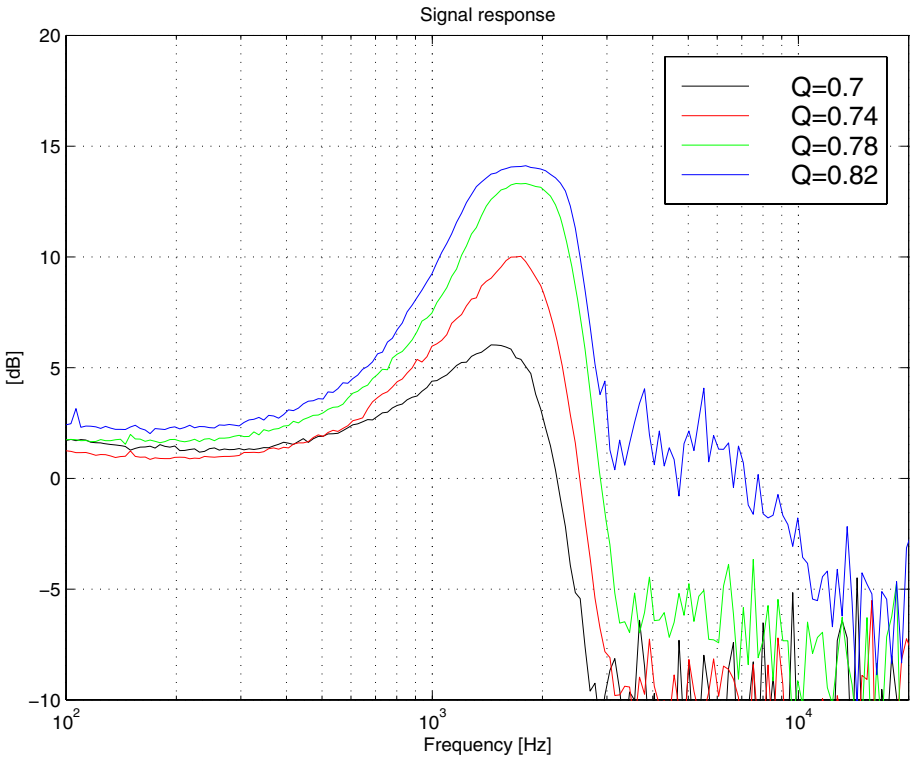


Figure 2.24: Measured cochlea response on output 50 with varying Q.

Finally figure 2.24 show a scan for four different Q settings. To keep the amplifiers out of saturation while a strong resonant peak is created, the input amplitude was set to 100mV limiting the lower gain resolution to approximately -15dB. Although adjustment of Q affects the signal response as desired, the real absolute value is higher than the computed, as discussed in section 2.4. But most of the observed strong resonance stems from pseudo resonance creating a much stronger resonant peak than one single section. Note that the cascade is stable for the highest Q setting, even though the output saturates at approximately 15dB.

2.8.3 Second order filter section

All measurements were executed on the first filter section in the filter cascade. They were controlled from a MATLAB script, but for adjustment of Q and w_o to a desired value, g_m of the amplifiers must be known for a given bias voltage. To be able to compute g_m from the bias voltage V_b, the relation

$$g_m = e^{116.7404 - 30.3397 V_b} \tag{2.15}$$

has been deduced from the simulation results in figure 2.16 in section 2.5. See section A.3 in appendix A. This relation can easily be transformed to

$$V_b = \frac{116.7404 - \ln(g_m)}{30.3397} \tag{2.16}$$

The equation can be used to find the appropriate bias voltages for a given Q and w_o after the desired g_m is computed by equation 2.12 and 2.13. In the following measurements C_1 is assumed to be equal to C_2 and estimated to $1.5pF$ from the layout including major parasitic capacitance. In all measurements the voltages are computed using a MATLAB function implementing the formulas above, and applied with a Keithley 213 through a GPIB interface.

Frequency response

To determine the frequency behavior of the second order section, a frequency scan has been performed for different resonance frequency settings.

The results in figure 2.25 are in good correspondence with the earlier Spectre simulation although the measurements evidently show noticeable damping for all signals. Although the quality of the measured curves is somewhat poor, a slight increase in damping can be seen particularly for the three upper resonant frequencies when the input frequency is approximately 100Hz. This result is in good correspondence with the prediction from the MATLAB simulation. Even compression of the actual w_o is evident as predicted by the same simulation. Note the almost perfect cross of the curve for the undamped resonant frequency $f_o = 10kHz$ at $0dB$.

In support of third order effects is an evident roll off of more than $40dB/dec$ which is expected of a plain second order section, and the fact that the Q factor is slightly larger at low frequencies.

But it appears to be an unbalance in the total filter capacitance as the peak value is larger than $4dB$ instead of about $1.25dB$ which is expected for a second order section when $Q = 1$.

Q factor

In the measured results in figure 2.26 the Q factor has been adjusted according to the simplified formulas 2.12 and 2.13.

From the figure the Q factor seem to vary as desired with change in the bias currents, although the absolute Q value is larger than the expected value if the capacitances are of equal size. This might indicate that the value of the capacitances is not exactly equal, which is not surprising considering the relatively large and unpredictable capacitive loading from the input transistor wells of the WLR amplifiers.

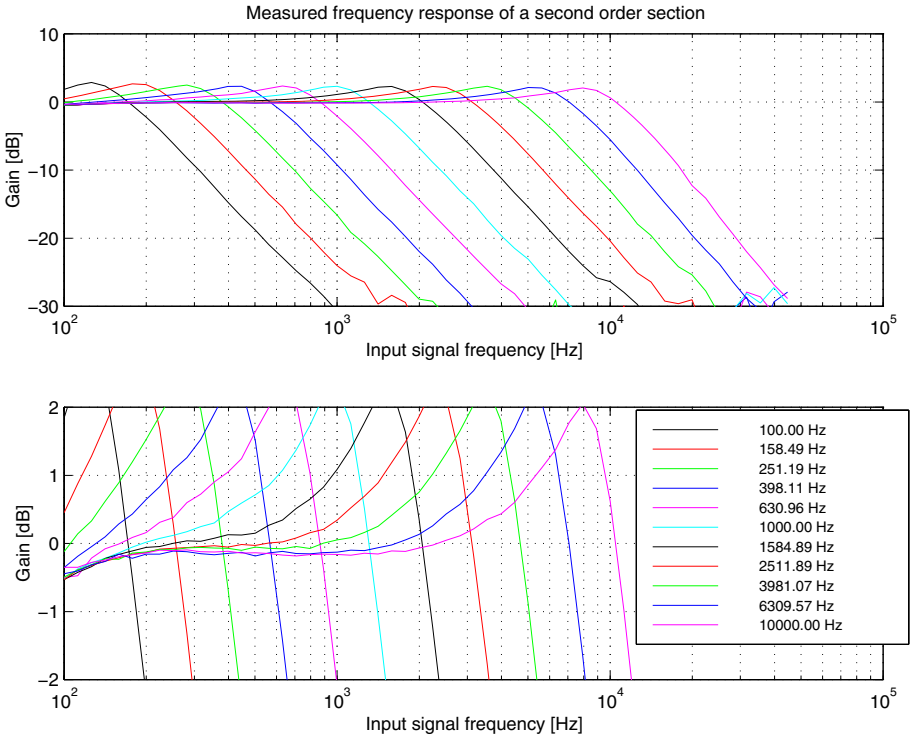


Figure 2.25: Measured frequency response of the second order section with bias voltages varied to adjust the resonance frequency (Q=1). The lower plot is the same results with rescaled y-axis to show the pass band damping.

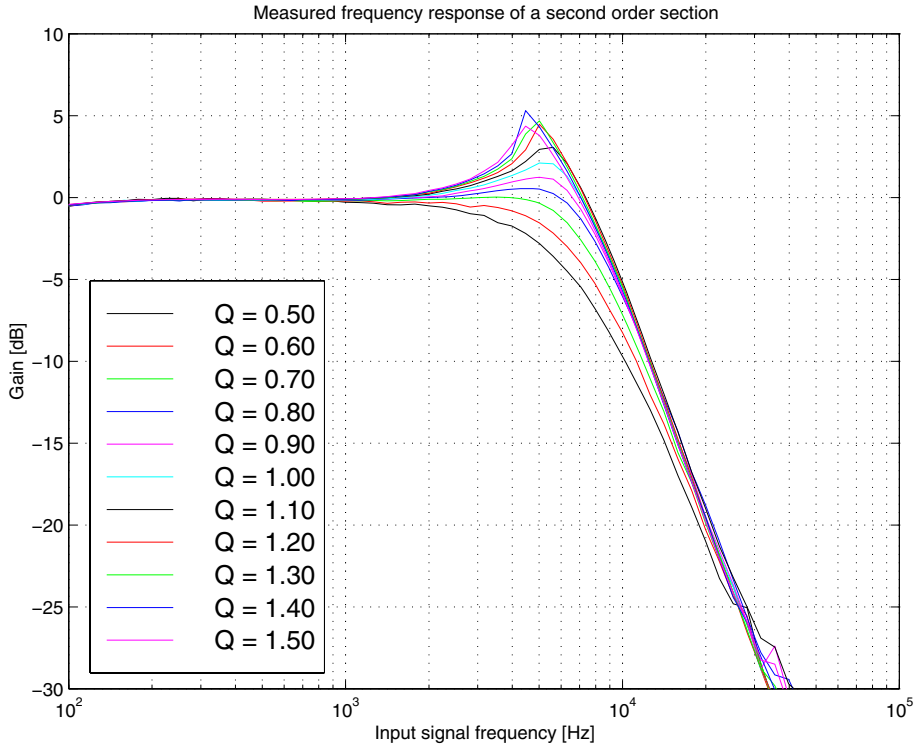


Figure 2.26: Measured frequency response of the second order section with bias voltages varied to adjust the Q factor. The undamped resonance frequency f_o is 6.31kHz.

2.8.4 Offset correction circuit

The actual improvement in number of functional stages with the OCR circuit enabled is indicated by the following measurements. They also show if the circuit adds significant noise or distortion to the output signal.

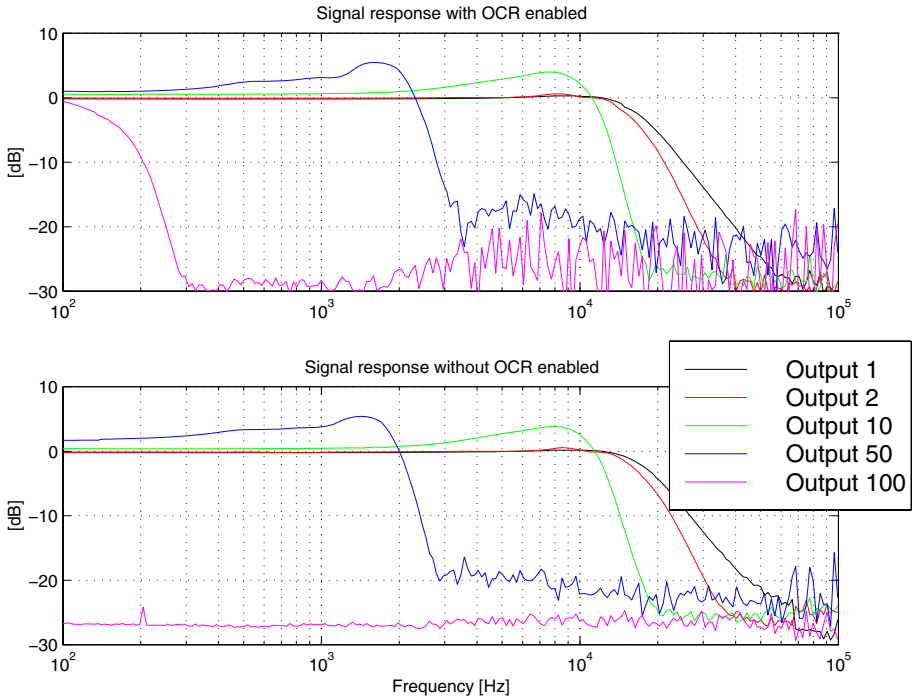


Figure 2.27: Plots showing the response of the cochlea cascade first with the OCR circuit enabled, and then in the lower plot with the OCR circuits disabled.

Measurements on the complete cochlea cascade in figure 2.27 show the frequency response with and without the OCR circuit enabled. From the first figure it is clear that the OCR improves the filter cascade as signal is evident even at output 100 when the circuit is enabled. No significant impact is evident on any of the other outputs indicating that they are operating with satisfactory operating range even without the OCR circuit.

To determine if the OCR circuit distorts the signal of a single filter section, the first filter cascade output has been monitored while the input frequency was scanned.

Figure 2.28 show THD and SNR for measured results with the OCR circuit enabled in the black curves, and disabled in the red. The results are practically identical with and without the OCR circuit. From these results it can be concluded that the OCR circuit does not contribute with any significant harmonic distortion components.

When measuring the harmonic distortion of the complete filter cascade up to output fifty, no significant difference can be seen for the distortion with and without the OCR

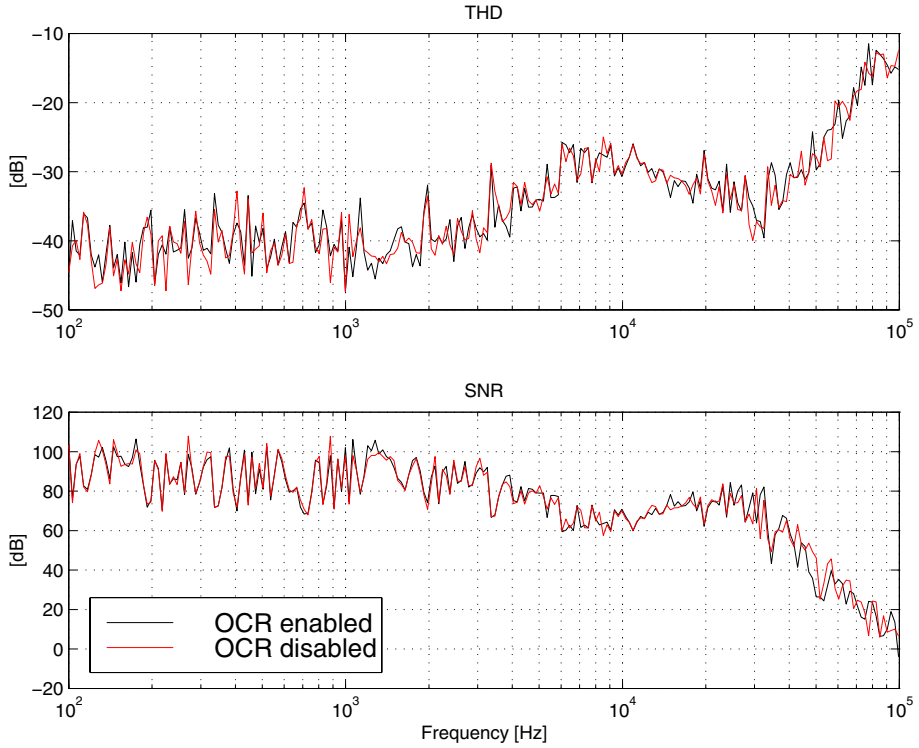


Figure 2.28: THD and SNR plots for measured results on a SOS section with and without the OCR enabled.

Output	OCR enabled	OCR disabled
1	-26.3 dB	-26.1 dB
2	-24.7 dB	-24.6 dB
10	-18.2 dB	-18.4 dB
50	-23.0 dB	-22.5 dB
100	-7.2 dB	

Table 2.4: Measured THD at the best frequency (BF) of the measured outputs.

circuit, see table 2.4. The exception is the last output which is not working correctly when the OCR is disabled, and the signal is suppressed.

Thus the OCR is performing well, and is improving the number of usable stages in the cochlea cascade. With the novel implementation, no measurable noise is generated by the OCR circuit.

2.9 Conclusion

The automatic Q-adjustment has not been implemented, somewhat reducing the benefit of increased sound quality, particularly when the input sound pressure is either very low or very high. There might even be a reduction in the percepted sound quality if an important frequency component is buried in environmental noise.

The phase response at the best places appears to be within the measured results for a biological cochlea. Examining the phase response at output 50 reveals a delay of approximately 3.5 cycles at the resonant frequency, which is well within the three to five cycles of phase accumulation reported by Rhode [39].

As predicted, the linear analysis from section 2.4 proves to be correct and very usable to determine the important design parameters for the circuit. The damping is evidently small if the cascode transistors are enabled, but the results with disabled cascode transistors clearly show why it is important to analyze the circuit well before it is implemented. The results are even supported by the SNR measurements which clearly improve with enabled transistors.

The novel analysis of the OCR section has lead to an implementation generating minimum harmonic distortion. From the measured results it is shown that the OCR circuit is a necessary part of the cochlea cascade, and the implemented design proves to satisfy the demands of low distortion and satisfactory DC correction.

Some of the improvement is clearly due to improved design, where most of the effect probably is due to the improvement of the WLR amplifier. In the original design only 39 outputs were functioning with the OCR correction enabled. In this design 50 outputs are working without this correction. With enabled correction even output 100 is working, although the signal is significantly distorted.

The cochlea cascade is showing very promising results, and surpassing both the usable number of output channels, dynamic range and frequency range of the original design in [41]. The design is even capable of operating over the entire frequency range with one filter cascade, improving the filter phase response as one continuous cascade is sufficient instead of three cascades operating over different frequency ranges.

Overall the filter cascade is a promising candidate as signal processing unit of a cochlear implant.

Chapter 3

Voltage-to-current converter

3.1 Design

A conversion circuit is necessary to adapt the cochlea voltage mode outputs to current mode for the neuron circuits. Although transconductance circuits are relatively simple to implement, two major challenges directs the design. Again the large number of channels dictates that the space and current consumption must be small, demanding a compact low current design.

To convert voltage to current, a restive device must be coupled in series with the voltage source. In theory a passive resistor could be used, but since the conversion factor is in the order of one million, an unpractical large passive resistor would be necessary with the chosen process. It would in addition be troublesome to make the resistance adjustable for to implement adjustable the gain factor.

A more rigid solution consisting of an OpAmp with four resistive devices and a load impedance is proposed in [33], (eq. 10-116 and figure 10-48, p. 449). With this solution the input impedance is independently controlled by the input resistors while the conversion gain is controlled by the load impedance. Again the main limiting factor is the resistances as four resistors are required in addition to the load impedance. Even though a single transistor can replace the load resistor, there are four resistors which either must be implemented as plain space consuming passive devices, or as HRES devices [31] which would be complex, match poorly and have poor linearity.

A third option is to use a current conveyor [44]. A simple transconductance converter is implemented with a type II conveyor and a resistive device connected to the X input. The conversion factor is then given by the inverse of the resistance, still demanding a high resistive value. The major benefit of the conveyor solution to the OpAmp implementation is that there is only one resistor which is grounded in one end, only demanding a single MOS transistor with a voltage bias setting the resistive value. The main drawback of this solution is the current consumption. A typical class II design as reported in [4], is designed for above threshold operation with focus on parameters like bandwidth instead of current consumption.

A simpler solution is to use a transconductance amplifier, but then the output of the OTA will be connected to a current mirror keeping the output of the amplifier out of range. To solve this problem, a transistor operated as a resistive device has been added in series with a voltage follower connected OTA. This transistor converts the output

voltage of the follower to a current. The gate voltage is controlled by a voltage source circuit adjusting the effective resistance of the device, thus permitting an adjustable conversion factor.

3.2 Architecture

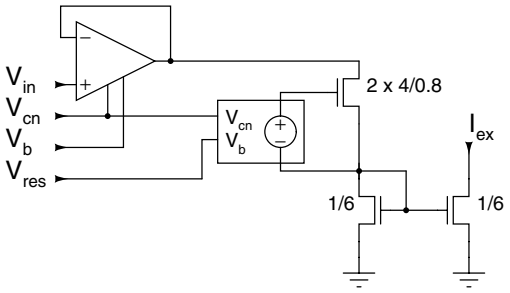
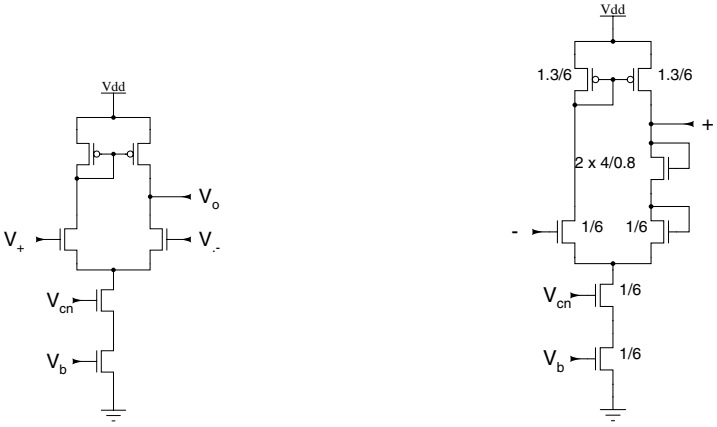


Figure 3.1: Schematics of the voltage-to-current converter. The box containing the voltage source indcates the voltage source circuit.



(a) OTA (b) HRES-voltage source

Figure 3.2: Figure a show the schematics of the OTA and figure b show the voltage source circuit used in the voltage-to-current converter.

In contrast to the circuit (HRES) described by Mead, the current flow is unilateral, requiring only one transistor biased to operate as a resistor.

The resistance is adjusted with the gate-source voltage, but because the source voltage is varying with the current through the transistor, a voltage source circuit [31] has been implemented, see figure 3.2. The voltage source is a simple voltage follower with

an additional diode connected at the output, adding a V_t offset to the output voltage. Because the voltage drop across the diode connected transistor is current dependent, the offset of the output voltage can be adjusted slightly with the bias voltage V_b. If the current is increased, the voltage-drop increases, increasing the voltage offset. In effect, the circuit functions as a voltage follower with an offset dependent of the current through the offset diode.

The OTA design, also shown in figure 3.2, is identical to the OTA used in the LP-filter in the SOS-sections. It is used to reduce the number of different sub cells, even though the cascade transistor can be removed as the linearity of the OTA is unimportant when it is operated as a voltage follower.

3.3 Simulations

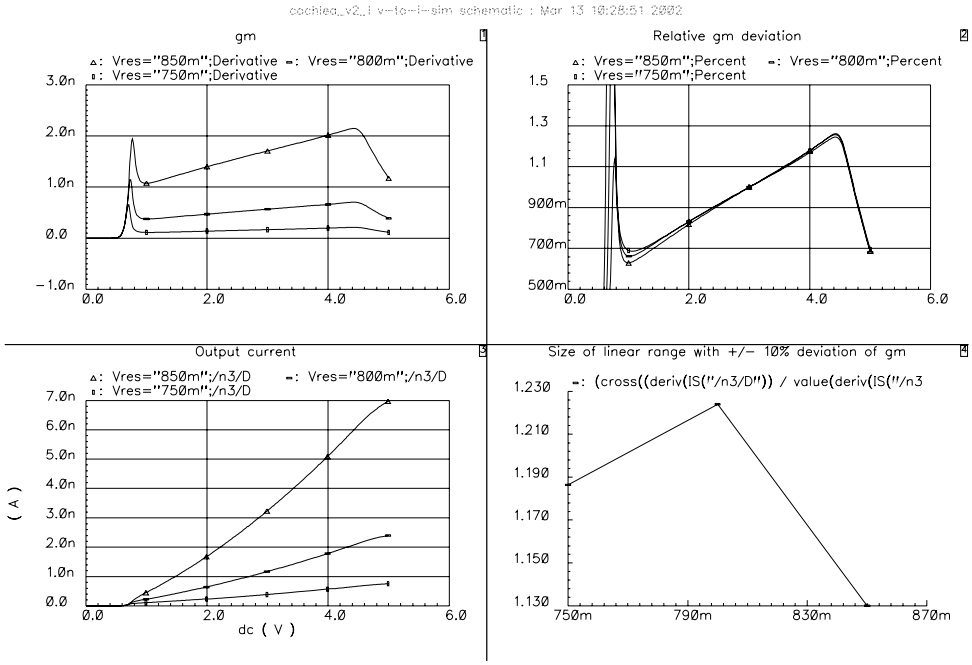


Figure 3.3: Simulation results on the voltage-to-current converter. The simulation has been rerun with three different bias voltages on the voltage source circuit. The upper left panel show g_m computed as the derivative of the output current, the output current is shown in the bottom left panel, the relative g_m deviation at 3.0V is shown in the upper right panel, and finally the lower right panel show computed linear range for g_m with less than 10% deviation for the three bias voltages.

Linearity and range simulations are shown in figure 3.3. Typical operation point is 3.0V, and typical amplitude is less than 1V. The linear range computations show satisfying range (less than 10% deviation of g_m) is larger than 1.13V, although this is

deependent of the bias voltage. If the input range is 1V with 3.5V used as operating point, the distortion will be approximately ±9%.

3.4 Measurements

Due to pad limitations, the output of the current mirror was not made externally accessible. Measurements have to be performed on the input node of the current mirror, but the low current make the circuit very noise sensitive, making accurate measurements very difficult.

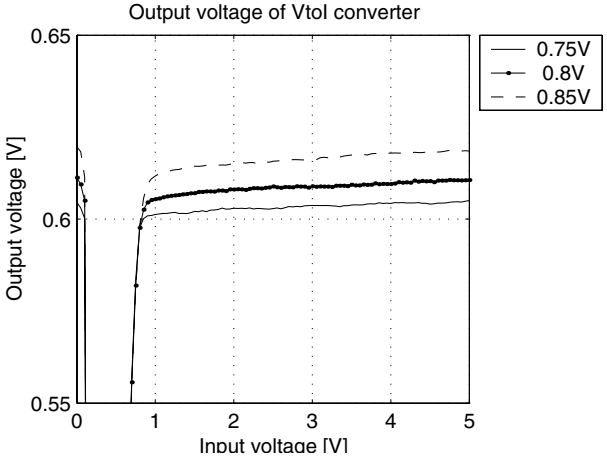


Figure 3.4: Voltage measurements on the input node of output current mirror of the voltage-to-current converter for varying V_{in} voltage. V_{res} is varied for the three different curves.

Both voltage and current measurements on the output have nonetheless been performed, although the resulting numbers are far from accurate. In figure 3.4 voltage measurements indicate correct dependence of the V_{res} bias voltage as the output voltage increase as the V_{res} voltage is increased.

Measurements showing output current dependence of input voltage is shown in figure 3.5. Here the input node current of the current mirror is measured when the output node is tied to ground, effectively bypassing the current mirror. Note when the source of the resistive transistor is grounded, the voltage source circuit operates out of range. This is evident as the measured slope of the output current is much steeper than for the results from the simulations, indicating increased gate-source voltage. If the response of the following neuron is observed when the current mirror input is at desired operating voltage, the voltage source circuit appears to be operating closer to the simulated results.

Also when observing this response, a large variation of the different voltage-to-current converter responses are evident. This discrepancy is probably caused by poor matching of transistors in the voltage source circuit which might be operated in sub-threshold.

This creates large mismatches in the output currents even for small process mismatch, which has not been tested with the circuit simulations.

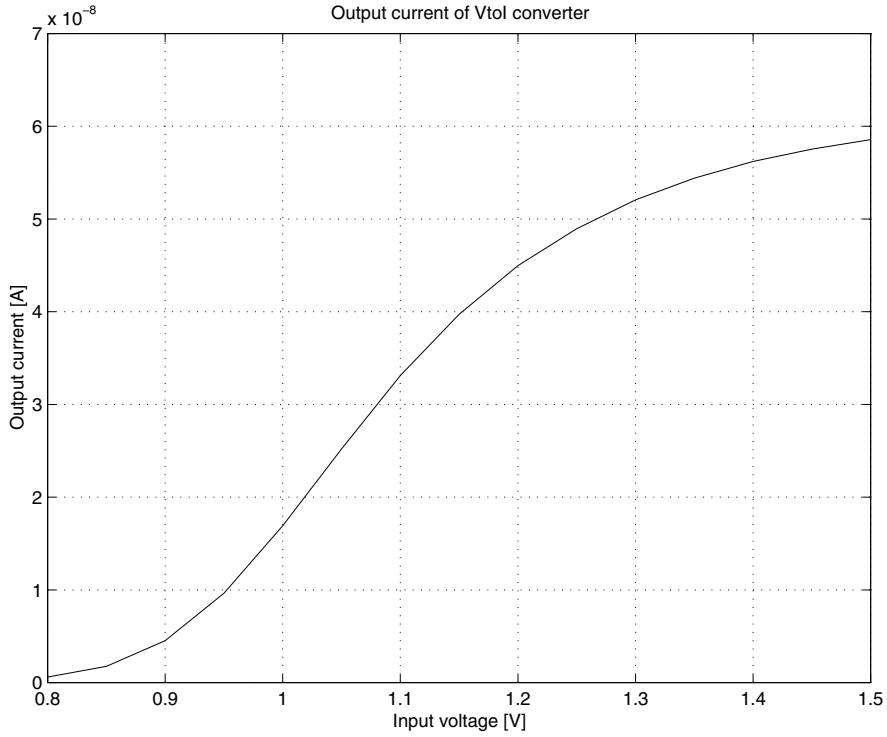


Figure 3.5: Current on the input node of output current mirror of the voltage-to-current converter. Offset is set to 0V, short circuiting the current mirror. The input voltage V_{in} is varied.

3.5 Conclusion

According to simulations, the circuit operates satisfactory for the intended application. But unfortunately the simulations did not reveal the mismatch problems of the real circuit, creating a far from acceptable difference of the converters of the different cochlea channels.

If instead an OTA with a PMOS differential pair is used, the negative output range of the amplifier is extended. This could be sufficient for a direct connection to the current mirror of the following neuron, but it might even be necessary to add a diode coupled transistor in series with the input NMOS transistor of the following neuron circuit to ascertain the OTA is within operating range. Thus the circuit can be reduced to a simple OTA and a diode connected transistor, generally saving both space and power while the matching should be improved.

Because the transconductance of the OTA is dependent of the current from the biasing current source, the overall transconductance of the converter is easily adjusted by the gate voltage of the biasing transistor. Finally the differential input of the OTA must be converted to single ended by biasing the negative input node to the DC level of the input voltage. This level could probably be extracted from the SOS offset correction circuit, which extracts the internal DC level of the SOS section, see section 2.6.

Chapter 4

Signal mixer

4.1 Neuromorphic signal processing

Post-processing of the cochlea output is a significant part of the system. Like the cochlea cascade, biological systems are used as inspiration for this task as well. It appears that biological hearing use a large network of neurons for signal conditioning before the different sounds are interpreted by the brain. Although this thesis does by no means try to make an exact model of such a network, the thought of using a large neuromorphic network to obtain scalable, flexible and robust signal conditioning has inspired the chosen solution.

Traditionally it has been assumed that neural spike intensity is the key information carrying factor for neural communication. But this theory is not capable of explaining very fast reactions, like the extremely fast reaction of the fly. It is apparently able to react long before the neural system is able to gather the signal mean value. Likewise, the human eyes are receiving a tremendous amount of information to be interpreted, but the resulting neural activity is much lower than expected for mean rate coding [40].

The major shortcoming with mean rate coding is that the relative spacing between the spikes becomes irrelevant. Thus the mean rate encoding is not as efficient as one would expect of a system developed through thousands of years of evolution. The major remaining question is then how the full bandwidth can be exploited without complex circuitry? The answer can actually be found in the theory of conventional $\Sigma\Delta$ converters which possess properties capable of explaining complex phenomena of biological systems.

Recent theories suggest that the relative spatial location of the spiking neurons is playing an important role together with the sequence of pulse arrival at the different locations [40]. When viewed in time domain, $\Sigma\Delta$ encoding utilizes the actual patterns embedded in the data stream, not only the mean value. The relative bit patterns determine the signal frequency components, while the average value is proportional to the signal DC-level. This theory will be further elaborated upon in section 4.1.1.

If an asynchronous bit-stream is serialized through a delay-line, spatial locations will instantaneous reveal temporal information of the data. This can be used for detection of temporal patterns in the stream, but a similar trick might be used to create a high rate bit-stream exceeding the capability of a single neuron. If multiple neurons are allowed to parallel process the same input signal, they are interconnected to inform all other neurons of spiking events, they can theoretically behave like one super-neuron far

exceeding the maximum spike-rate of a single neuron. Indeed this is probably done by biological neural systems [40]. It has recently been proposed that this strategy might be applied by clocked $\Sigma\Delta$ converters [5].

There are two important reasons for having multiple neurons encoding the same signal. One benefit is fault tolerance and graceful degradation which gives a functioning system despite faults in subsystems, at the cost of slightly reduced performance. The other reason is that biological neurons have a low maximum spike rate, limiting the overall signal to noise ratio. Adams showed in [1] results remarkable similar to $\Sigma\Delta$ converters using a software model of a neuronal system with collaborating neurons. According to this theory, the noise shaping is spatially distributed across a neuron population to increase the dynamic range of the system. This is achieved with a weighted global inhibitory feedback over the entire neuron population, reducing the spiking probability of the neurons closest to the last spiking neuron. Thus in accordance with the discussion above, the relative placement of the spiking neurons is important.

In the following sections, this theory is used to implement a microelectronic system for signal processing in cochlear implants.

4.1.1 Neuromorphic modulation

Several sources [6, 1, 22, 11] indicate relationship between $\Sigma\Delta$ converters and neurons. In [6] a relationship is suggested but not shown mathematically. In [1, 22] the concept of spatial noise-shaping is presented, where multiple neuron-models are interconnected with inhibitory connections to improve SNR. Finally in [11] in part II the direct relationship of neurons and traditional $\Sigma\Delta$ converters and FDSM converters are shown. In this section an alternative proof is introduced using a block-level model of a transistor neuron that shows it can be understood by theories developed for conventional $\Sigma\Delta$ converters. In addition to previous known material, the difference between a biological neuron and a $\Sigma\Delta$ converter will be discussed.

The purpose of the neuron is to convert the signals from continuous time and value to continuous time but quantized value domain. This is done by self-timed single level or monopolar converters. By single level is meant converters that generate an output that has two states; either an output is generated, or the converter is in the resting state and no output is generated at all. Note that conventional converters will have symmetrical output around zero, thus they will be able to generate both positive and negative output signals, in other words a bipolar output signal. Note the neuron model implementation is based on work done in [15], and the implementation is described in section 4.2.

When modeled at system level, the model in figure 4.1 applies. Many readers will recognize this as a model of a conventional $\Sigma\Delta$ converter that has been made completely asynchronous. In addition to what is shown in the figure the quantizer is monopolar and self-timed. Thus because the neurons can be modeled like traditional $\Sigma\Delta$ converters, they should even possess noise-shaping. In addition they will be able to over-sample the input signal at a rate determined by the maximum neuron speed, and they have an inherent activity adaptation. If no input signal is applied, the circuit will cease to generate output and consume very little power.

The signal transfer and noise transfer functions (STF and NTF) are straight forward

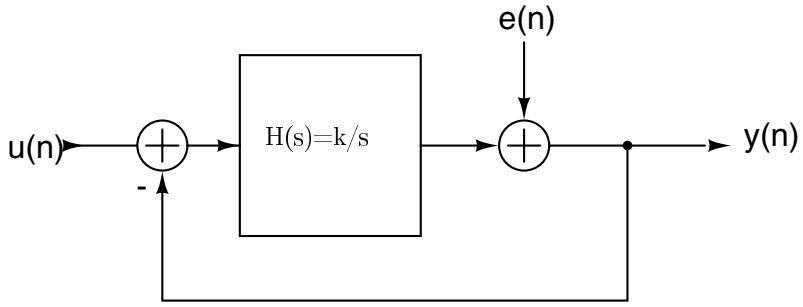


Figure 4.1: System-level model of the neuron.

to deduce, and the functions will entirely be developed for the S-plane because even the quantizers are continuous in time in the sense that they are able to generate an output at any time instance. From the equation

$$Y(s) = STF(s)U(s) + NTF(s)E(s) \tag{4.1}$$

STF is found as

$$STF(s) = \frac{Y(s)}{U(s)} = \frac{H(s)}{1 + H(s)} \tag{4.2}$$

$E(s) = 0$, and NTF is found as

$$NTF(s) = \frac{Y(s)}{E(s)} = \frac{1}{1 + H(s)} \tag{4.3}$$

if $U(s) = 0$.

For the neuron the integrator-function is $H(s) = \alpha/s$, where α is a tunable constant. Not surprisingly these equations are identical to conventional $\Sigma\Delta$ equations except that they model a continuous time system. Note that in these equations an optional leakage current dependent of the integrator charge is left out. This signal will permit compression of strong input signals as it increase with increasing integrator charge, but it should not significantly alter the primary function of the neuron shown by these equations.

Another very interesting property of the neuron is reduced sensitivity to aliasing. First the sampling is moved to the quantizer, like continuous time modulators [37]. This will by itself dramatically dampen high frequencies by the integrator gain. In addition, since the quantizer is self-timed, input signals with a frequency over the maximum neuron pulse-frequency will tend to generate a weak and relatively random output pulse sequence instead of a strong stationary in-band signal.

Utilization of a monopolar bit-stream implies that when the neuron fires, it will only subtract charge from the integrator. Since it is not capable of producing negative output pulses, there is no need to add charge. This simplifies both the DAC and the quantizer, and most important, it makes the system inherently stable. The only criterion for stability becomes that the maximum integral charge of the DAC must be larger than

that of the input-signal. If in addition a weak negative DC signal is applied at the input, the converter will reach absolute resting state and save power if no other input signal is applied. This automatically scales the consumed power with the strength of the input signal, removing the need for power-down state for very low-power systems.

Please see section 4.4 for simulations and measurements on different combinations of neurons.

4.1.2 Mixer concept

The implant must be adaptable to the wearer because there will be large variation in the residual hearing of the potential prosthesis subjects. It is important that this adaptation is quick, simple, and stable in the sense that the profile should be remembered many years without reprogramming. Further the power consumption must be low to achieve an acceptable battery life span. This is achieved using an architecture based upon the previous discussion by combination of neuron circuits converting the analog outputs of the cochlea to digital domain, and a mixer routing the digital signals through a network of gates. The novel proposed design has the benefit of low complexity and good scalability, both very useful if implementing high performance matrixes.

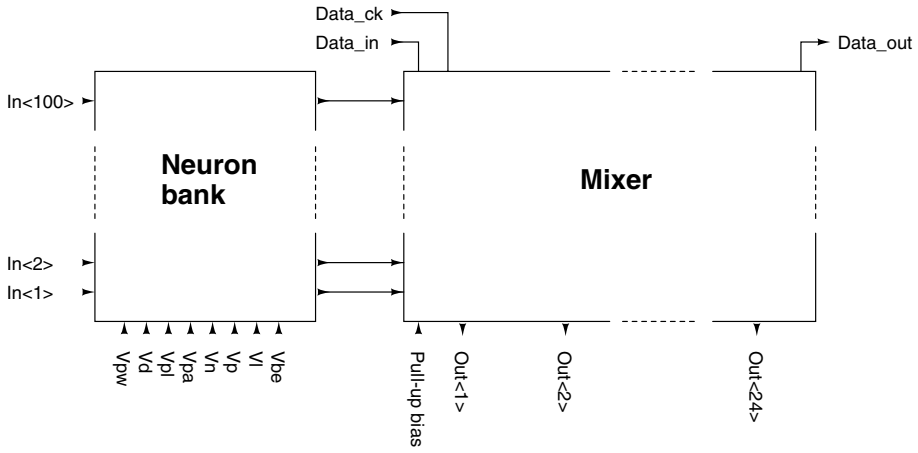


Figure 4.2: Block diagram of neuron-bank and mixer. See the text for explanation of the bias voltages. The neuron block is explained in detail in section 4.2, and the mixer in section 4.3.

In figure 4.2 the two major blocks of the circuit is shown. First the analog signal is converted to spike domain by a bank of neuromorphic spike-modulators. Then the output of these modulators is mixed by a mixer-matrix to adjust the frequency content and sound level of the different implant channels.

In [30, 11] included in the article section, it is shown that a single model neuron exhibits noise-shaping properties similar to what is found in $\Sigma\Delta$ converters. But for the intended application as a mixer for audio components, the dynamic range of a single neuron is insufficient. To improve the performance, a simple weighted global feedback scheme is implemented. With charge subtraction from the integrating capacitor of the

neighbors of a spiking neuron, the overall spiking probability is reduced. Thus if the integrating capacitor of all involved neurons is regarded as a single integrator, the global inhibitory action corresponds to the discharge of the capacitor of a single neuron when it produces a spike.

To increase the dynamic range of the neuron of a single cochlea output, it is possible to connect multiple neurons to the same output. This might however not be necessary because the frequency components of the cochlea outputs are grossly overlapping, hence the same signal will be present on multiple outputs. Thus it is possible to save both space and power by connection of all neurons to the same global inhibitory feedback.

On the implemented test-chip a solution with one neuron per output is chosen. To compensate for the relatively low single neuron dynamic range, the maximum spike rate can be increased beyond the maximum spike rate found in biological neurons. In the measurements for article [30] in part II, a spike rate of approximately 5kHz was chosen. Thus by utilizing the electrical properties of silicon a much better performance is possible using a single neuron.

By extrapolation of the SNR found with MATLAB simulations in [30] to one hundred neurons, the collective expected SNR of the complete neuron block is found to be 66dB. This is roughly equal to the SNR offered by commercially available digital cochlear implant processors. This figure can be further improved by increased spike rate at the expense of higher power consumption, if desirable.

Obviously, the dynamic range available for a single cochlea output will be much lower than the total cumulative SNR. But because the frequency content of the outputs is overlapping, the SNR for a single pure tone will be larger than the SNR of a single neuron. With the example settings the SNR of a single neuron will be approximately 20dB, and the SNR of a single tone will be larger than this figure, depending on the number of active neurons. Particularly the center audible frequencies will be processed by the largest number of neurons, forming a similar response as found with audiology tests showing that the ear is clearly most aware of these frequencies.

4.1.3 Signal mixing

As mentioned in chapter 1, the implant must be adaptable to different patients, and preferably have separate settings for different listening conditions. For easy configuration, both from a computer or from local program storage, the implant should be digitally configurable. The computer programming is needed when the implant is adapted to the patient, and the local storage is used for user selectable settings for different listening conditions. Because reprogramming of the device is relatively seldom executed, speed is not a high priority. It is therefore not a problem if the mixer must be completely reprogrammed each time a setting is adjusted. The programming interface might be serial, as its speed is sufficient for the application, as shown in section 4.3.2. With serial interface only two pins are needed for reprogramming, clock and data in, and a standard external PROM could easily be used for storage of the settings.

The patient adaptation consists both of frequency band adjustment and signal amplitude adjustment of the different electrodes. It can either be done directly by mixing the analog outputs of the cochlea, or the signals can first be converted to pulse domain before they are mixed.

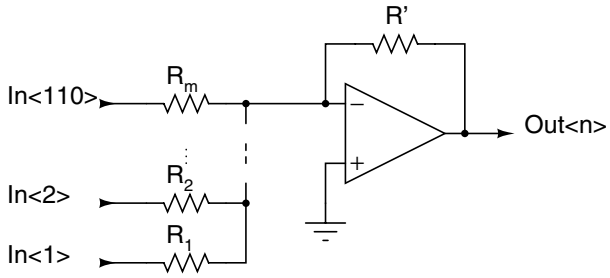


Figure 4.3: Typical implementation of an analog mixer.

If the first solution is applied, it will typically be implemented with one OpAmp-mixer [33] for each output channel, see figure 4.3. With this architecture, m times n input resistors are needed, where m is the number of input channels and n is the number of output channels. The resistors must be large to suppress interaction of the input channels and create high impedance inputs. High resistance passive resistors would consume much area, and active resistors in the form of transistors yield a matching problem and require complex biasing circuitry for each resistor [31]. Furthermore, the active resistors would have a limited linear rage, limiting the dynamic range of the mixer.

On the other hand, digital mixing require m A/D converters, whereas analog mixing only require n, but the mixing cells are digital and require much less space than the resistors of the analog version as only one minimum sized transistor together with a global pull up transistor is needed to mix the digital signals.

Assuming that the signal is spike coded with a stochastic distribution, signal mixing can be performed with summation of the pulses [36] from desired neuron outputs to form the stimuli of a cochlea electrode. Addition of multiple pulse streams generated by the same signal will increase the signal power, and addition of pulse streams generated by different frequency components will generate a pulse-stream representing the composite signal. Hence adjustment of signal power is performed by addition of a desirable amount of outputs from neurons stimulated by a similar signal, and the frequency content is adjusted by selection of neurons stimulated by the desired frequency content. Because the implemented chip has only one neuron for each cochlea output, the power adjustment is limited by the overlapping of the resonant peaks of the cochlea.

The drawback of the scheme is pulse collisions. Because the pulses are assumed to have a stochastic distribution, there exists a risk of pulse collisions. With the proposed scheme this will easily lead to a pulse loss giving signal distortion. The solution to this problem is a global feedback which reduces the spiking probability of all neurons right after a spike has been generated. With proper weighting the collision probability can be reduced to an ignorable level.

Programming of the digital mixer is much simpler than for the analog version. The digital bits might be set directly from the data stream without any transformation, whereas the analog version needs at least one global D/A converter, an analog multiplexer and local temporary analog memory for all the input resistors.

4.2 Neuron bank design

The neurons has been developed from the design described in [15]. The axonal function has been separated to permit an adjustable pulse delay to model the pulse traveling along the axon. To support external inhibitory feedback from neighboring neurons an inhibitory input has been added. Finally the neuron buffers have been altered to reduce power consumption. Figure 4.4 show a block diagram of the complete design.

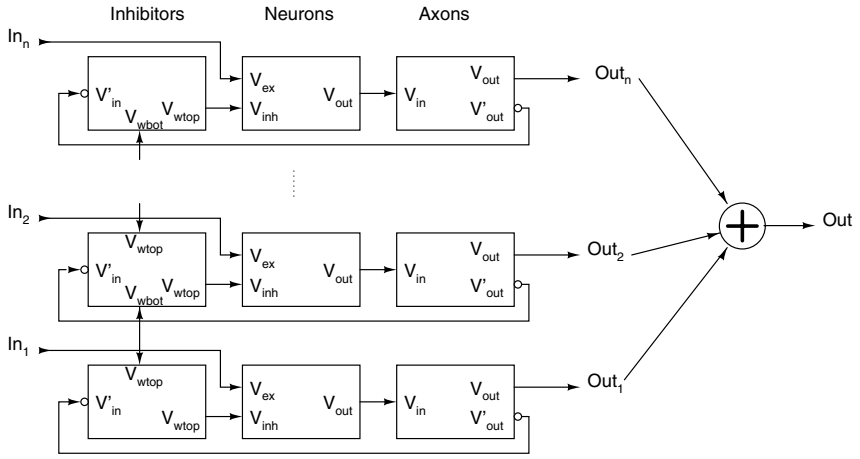


Figure 4.4: Block diagram showing the design principles of complete slices of the neuron system.

4.2.1 Self resetting neuron

In figure 4.5 the schematics of the neuron core is shown with the adaptation circuit grayed. Assuming that the sources of the input current mirror are at the same potential and V_{inh} is constant, the pulse rate at V_{out} is directly proportional to the input current. The pulse width is set by V_{be} while input V_{inh} is used for inhibitory action. It is activated by amplitude modulated pulses from the inhibitory circuit described later in section 4.2.2, and subtracts charge proportional to the amplitude of the pulse from the neuron capacitors.

The temporal adaptation circuit is identical to the circuit described in [15], except that a buffer from V_{out} to the input of the adaptation circuit has been removed to save power and space. This buffer has no functional importance except to increase the gain of the feedback loop created by the adaptation circuit. A sufficient gain is already present in the loop due to high transistor gain.

The scaling loop marked in gray serves to adapt the pulse stream to the intensity of the input signal. If the integral sum of the pulse stream is large, the gain in the input current mirror is scaled down to reduce the neurons sensitivity of the input signal. The integrating function is implemented by the feedback capacitor in the grayed box and the two NMOS transistors at the source of the input current mirror adjusting its gain.

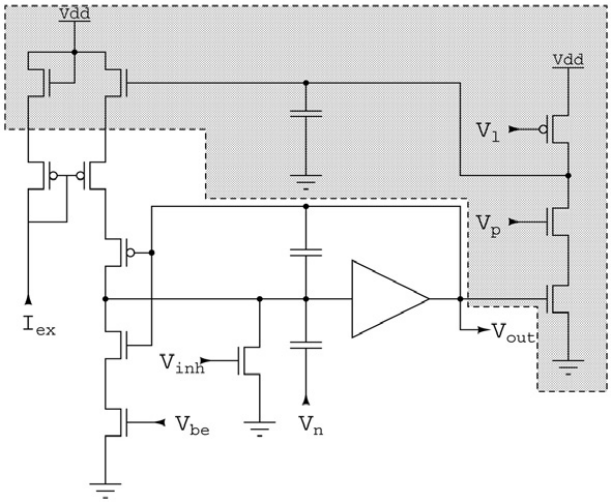


Figure 4.5: Schematics showing the design of the neuron core.

The capacitor is discharged with a fixed value for each output spike, controlled by V_p, and constantly charged by a fixed current controlled by V_l resetting the scaling when the signal intensity is reduced.

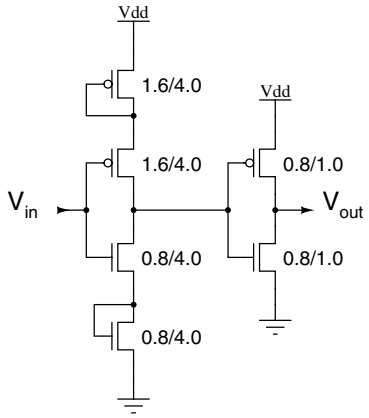


Figure 4.6: Schematics showing the design of the low power neuron buffer.

Low power buffer

In [16] a low power buffer is suggested, but this requires two additional bias voltages. Because the circuit already contains a large number of bias voltages, a bias-less low power buffer was designed.

The neuron buffer, shown in figure 4.6, has been designed to reduce the current consumption and noise generation of the neuron. By reducing the current consumption

of the first stage of the buffer with source degeneration, the current consumption has been greatly reduced as this stage is in an intermediate state for a prolonged period when the capacitor voltage is close to $V_{dd}/2$. The second stage is in this intermediate stage for a shorter period as the pulse is already partly shaped. This buffer hence consists of a traditional inverter with high gain and its main purpose is to create square output pulses.

Figure 4.7 show the current consumption of the new buffer compared with a simple two transistor buffer when the neuron spikes at approximately 80kHz[1]. The mean current consumption for the two simulations are 16.2μA for the standard buffer, and 0.8μA for the low power version.

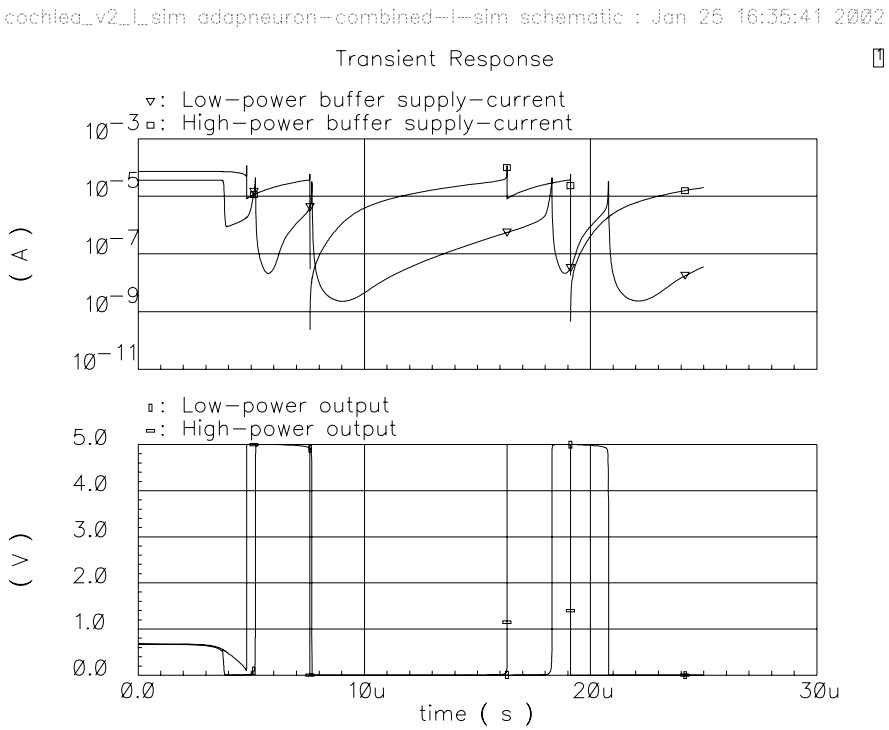


Figure 4.7: Simulated results showing the current consumption of the low power buffer and a traditional buffer when the neuron spikes. Particularly the region between the spikes is interesting because of the close to switching threshold input voltage. In this region the current consumption is roughly two magnitudes lower for the low-power buffer when it is compared with the standard buffer.

[1]The neuron with the high power buffer spikes at 87kHz, while the low power version spikes at 76kHz. This gives a 13% difference in spiking rate, and is of little importance to the huge difference in simulated current consumption.

Neuron-speed

A biological neuron in the auditory path will typically lock to the signal frequency and submit maximum one pulse per signal period. But the very limited dynamic range of a single neuron is improved with redundancy by parallel coupling of several neurons until a satisfying range is reached [22, 30]. This strategy is implemented in the cochlea cascade by grouping several outputs which are close in frequency range. To further increase the dynamic range, the neurons can be adjusted to permit a maximum pulse frequency which is higher than the period of the input signal.

Although most transistor ratios have been kept as they are in the original design, the input transistor has been modified to permit a moderate spike rate in the operating regime. The capacitors and all transistor sizes have been reduced to increase the mismatch effect to add a random effect reducing the probability of pulse collisions.

4.2.2 Inhibitory circuit

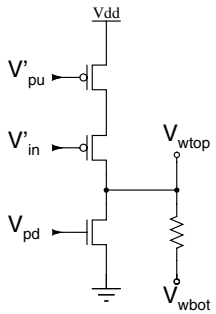


Figure 4.8: Schematics showing the design of the inhibition cells.

The novel inhibitor circuit, shown in figure 4.8, are multiplied to create a resistive network forming a distance from spiking neuron dependent weighting mechanism for inhibitory feedback to all neurons. In the work of Adams [1] the term triangulation is used because the scaling is linearly decreasing from the source. This term is adopted here although the scaling is actually exponentially decreasing.

The pull-down transistors (V_{pd}) set the slope of the triangulation, and pull up transistors (V'_{pu}) adjust the maximum amount of charge subtracted from any neuron. When V'_{in} is pulled low by an axonal output, the V_{wtop} voltage is decided by the current-ratio of the pull-up and pull-down transistors of the current inhibition cell, and of the V_{wtop} and V_{wbot} voltages of the neighboring cells. To close the inhibition loop the V_{wtop} voltage is connected to the inhibitory voltage (V_{inh}) of the current neuron, causing a spike to subtract a charge inversely proportional to the distance from the spiking neuron(s) during the whole period of the spike. Because self-resetting neurons are used, the spike width is constant, casing the amount of charge only to be distance dependent as desired.

To disable the triangulation, the pull-down bias voltage V_{pd} can be set to zero. This will cause all the neurons to receive the same negative charge. Alternatively the inhibitory action of the neighboring neurons can be completely shut off by setting V_{pd} to a sufficient high voltage.

4.2.3 Axon

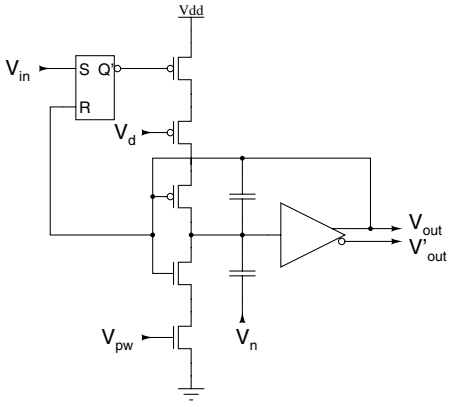


Figure 4.9: Schematics of the axon blocks.

Because an instantaneous discharge of the neuron capacitor causes the neuron to reset, only a short spike results if no axonal delay is added to the spikes. To prevent this, a separate axon circuit has been added on the neuron outputs. This circuit even permits examination of the importance of feedback path delay. In figure 4.9, the input SR-latch is set by an incoming pulse, and it charges the capacitors with a rate decided by V_d, hence setting the delay of the pulse. When the voltage of the capacitors reach the switching threshold, the output flips high and the capacitors starts discharging by the rate set with V_{pw}, which controls the pulse width.

Measurements

Function	Minimum	Maximum	Bias
Pulse width	59.5ns	17.0s	$V_{pulsevidth}$=5V/0.25V
Pulse delay	208ns	576ms	V_{delay}=0V/4.49V

Table 4.1: Measured delay and pulse width for the axon circuit. See the text for comments on the measure methods.

A few simple measurements have been executed to verify the function of the axon. In table 4.1 a summary of measured parameters is reported. Pulse width is measured directly with an oscilloscope when the bias voltage is adjusted. The pulse delay is measured as shortest delay between two consecutive pulses, even this with an oscilloscope.

Because the pulse delay is measured as the minimum delay between two consecutive pulses, and not the actual delay of the axon, the reported delay is inaccurate. Still, if the axon input is constantly high, it will generate a pulse-train with pulse space mainly decided by the axonal delay. For the minimum value, other parameters than the axonal delay become important for the pulse interval, but the measured delay is believed to give a fair indication of the actual delay.

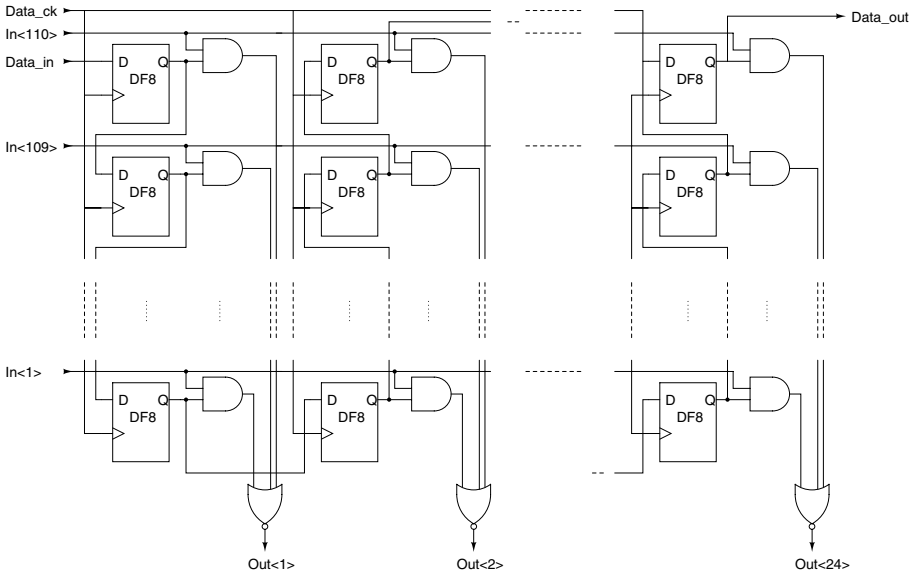


Figure 4.10: Structure of the implemented switch matrix.

4.3 Design of the switch matrix

Based on the previous discussion, a novel serially configurable mixer matrix is implemented to support fast and simple configuration of the device. Neuromorphic pulses are generated by a neuron bank, see section 4.2, with global inhibition. The neurons feed the matrix which distributes the pulses to 24 output channels. The current matrix has 110 input channels, where the first hundred is feed by the cochlea cascade, and the last ten are connected to external pads via neurons.

4.3.1 Architecture

As indicated earlier, the main features of the switch matrix are simple and fast configuration, and reasonable space consumption. For use in cochlear implants with specialized sound processing profiles for varying listening conditions, configuration programs must be downloadable from an external PROM. This is easily done with a serial interface which requires very little interface components as the data addressing is implicit by the position of the data in the bit stream.

The asynchronous pulse stream only quantizes the signal in amplitude, not in time. This is important for maximization of the SNR ratio, as shown in section 4.1.1. This is effectively ensured feeding the signal through a simple digital switch implemented by an AND gate, and summation the outputs of multiple gates to form the signal outputs. Each switch is set by a locally implemented flip-flop configured through the serial interface. Figure 4.10 show an overview of the implementation.

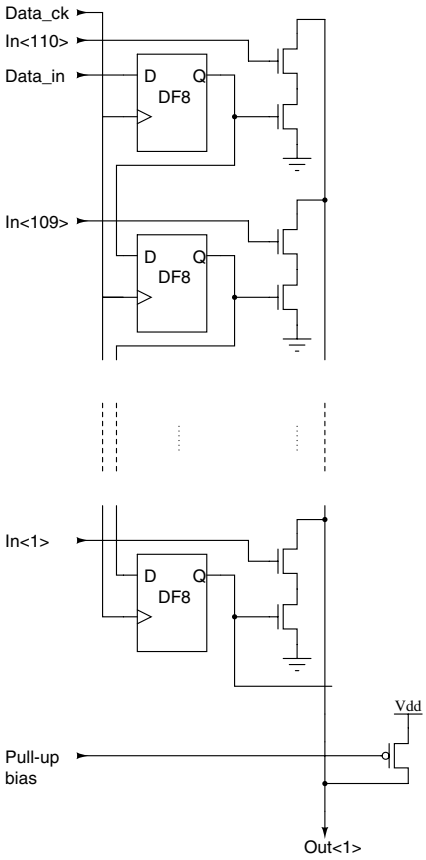


Figure 4.11: Details of a single column of the switch matrix.

4.3.2 Implementation

The switch cells shown in figure 4.11, are implemented with the DF8 [2] static flip flop from AMS HitKit 3.20 and two NMOS transistors. The transistors establish a two input AND gate together with a global pull up transistor biased to function as a resistive device. All outputs from the switch cell on the same row are NOR´ed through a wired NOR circuit biased with the same pull up transistor as the AND gate.

The clock signal has been routed in the same direction as the data signal, opposing established theories as this might cause setup-and-hold violations down the cascade. But it was none the less done to make sure that at least the first flip flops could be configured in case the clock signal was too distorted at the end of the cascade. But unfortunately this routing is the cause of setup-and-hold timing violations.

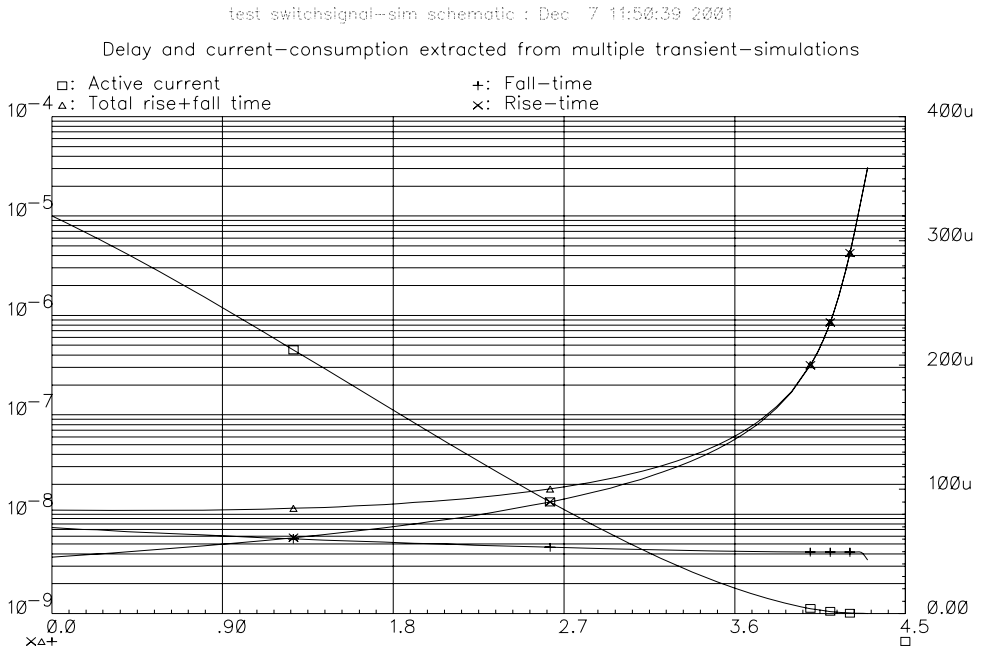


Figure 4.12: Simulated signal delay of the switch matrix. The results are extracted and computed from transient simulations.

Programming speed

Maximum programming speed is determined by the clock load of the DF8 and the driving capabilities of the clock pad because no distributed clock buffering is applied. By extrapolation of the pad (AMS, IB55 [2]) output slope with the output load generated by 24×110 flip flops, and a maximum clock frequency of 8MHz is found. This figure is pessimistic as it assumes full swing of the clock signal, which is strictly not necessary. Total programming time is 330μs or less, which means that the patient will probably only detect a sudden change in the perception of sound quality. To further reduce the probability of discomfort during reprogramming, additional AND-gates can be added to the outputs to switch them off during reprogramming to suppress any undesired output pulses.

Pulse frequency

The pull up transistor of the wired OR dominates the maximum pulse frequency, and the introduced delay is dependent of the bias voltage of the pull up transistor. The simulations of a single output shown in figure 4.12 and table 4.2, show a minimum delay of 10ns due to the rise- and fall-times of the wired OR output. Assuming that this delay is orders of magnitudes larger than all other delays in the system, a maximum pulse frequency of approximately 90MHz is achievable if approximately 280μA current consumption of the active output during the active pulse phase is tolerable. If 1MHz

Delay	Current consumption	Bias voltage
10.87ns	282.8μA	477.4mV
100.0ns	12.2μA	3.771V
1.00μs	1.589μA	4.116V

Table 4.2: Simulation results for a single output of the switch matrix and its output pad. Simulation results are computed from extracted results of transient simulations. The indicated current consumption is the maximum simulated current consumption when the output is active low.

maximum pulse rate is sufficient, current consumption of approximately 1.6μA can be expected when the output spikes.

Comparison of these results with simulations of the rise- and fall-times of the axon to switch matrix connection, confirm that the wired OR is the limiting factor in the signal speed of the switch matrix. Rise time is simulated to 9.511ns and fall time is 5.538ns, or a bit less than the minimum rise- and fall-times of the output stage.

4.3.3 Measured results

Testing of the chips reveal that it is impossible to configure the complete switch matrix correctly. In the output bit stream ones tend to disappear. To be able to get a one through the whole system, the bit stream must incorporate a block of about 200 ones. But fortunately, it is possible to configure 2-3 output rows correctly, making it possible to do most planned measurements. Both the number of configurable outputs and number of disappearing ones is chip dependent, but it is not dependent of the clock frequency.

4.3.4 Troubleshooting

It is likely that programming difficulties of the matrix stem from bad clock distribution. First, but only limiting the maximum clock frequency, is the large cumulative impedance through the system. Approximations indicate the routing contribute a resistive value up to 3.5kΩ and the inputs to the DFFs generates a maximum of 58pF, limiting the clock frequency to slightly more than 1MHz. A more serious problem is that the clock is routed in the same direction as the data signal. This causes a larger cumulative clock delay for some of the DFFs than for the data signal, increasing the probability of violations of setup-and-hold timings.

In figure 4.13 the schematics of the DFF with clock buffers and the implementation of the clocked inverters is shown. Besides the setup-and-hold requirements, it is important that the *cn* and *ci* signals switch approximately simultaneously. If the edges of the clock signal are too sluggish, the imprecision of the clock switching will significantly affect the flip-flop switching symmetry as shown later.

Error location

To determine the extent of the problem a programming test has been performed. First the matrix was preprogrammed with zeros, then a full sequence of zeros and finally a

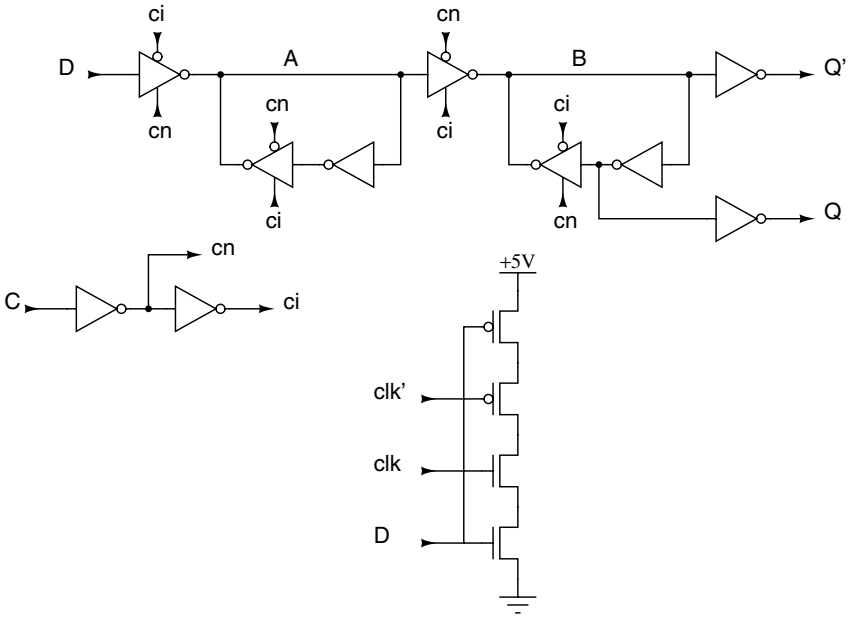


Figure 4.13: Implementation of the AMS DFFs. The transistor schematics show the interior of the clocked inverters and the clock buffers.

full sequence with ones were loaded and the output bits monitored. The returned vector was identical for all the zeroes, and then a vector of ones where the last 205 positions were replaced with zeroes followed. See table 4.3 for a schematic overview of the stimuli and monitored response. Note that the output vector is generated by the vector stored in the flip-flops; hence the expected pattern is equal to the previously loaded vector.

This result shows that ones easily skips clock cycles somewhere down the cascade as the trailing zeroes used for reading out the last vector appeared 205 clock cycles too early. But the exact error position is impossible to pinpoint. It is likely that the problem mainly arises at the end of the cascade, as this is where the clock signal must be most distorted, or in some of the junctions between the halves of the matrix. The routing between the halves add much resistance to the clock line and this might be sufficient to generate a significantly larger delay on the clock signal than the data signal.

Sequence	Input vector	Expected output	Measured output
1	$[0_1 \ldots 0_{2680}]$	$[-]$	$[-]$
2	$[1_1 \ldots 1_{2680}]$	$[0_1 \ldots 0_{2680}]$	$[0_1 \ldots 0_{2680}]$
3	$[0_1 \ldots 0_{2680}]$	$[1_1 \ldots 1_{2680}]$	$[1_1 \ldots 1_{2475}, 0_{2476} \ldots 0_{2680}]$

Table 4.3: Actual programming pattern, expected response during the next 2680 clock ticks and the actual measured response. The subscripts indicate the bit order in the data stream.

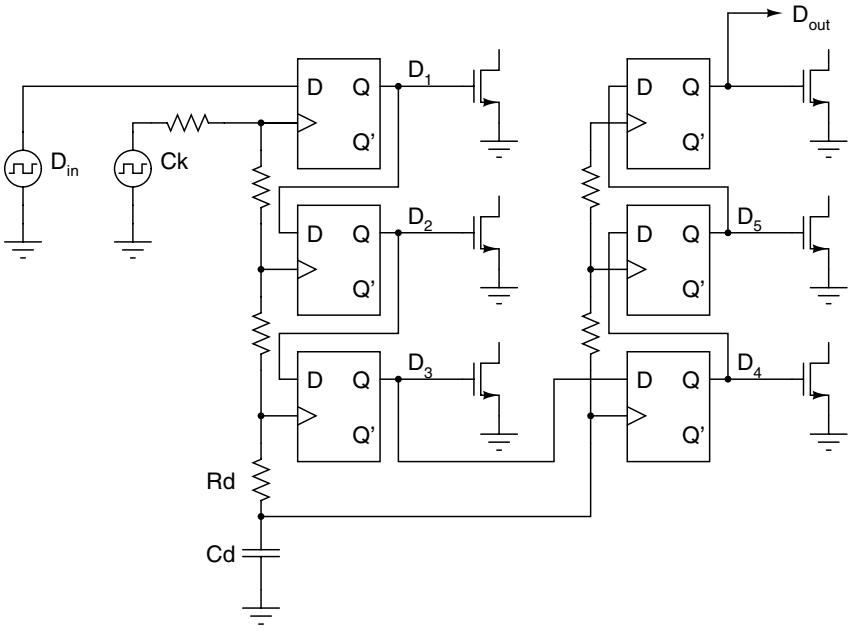


Figure 4.14: Simplified model of the scan chain with typical impedance values. *Rd* and *Cd* which are increased to simulate the large junction impedances of the two halves of the layout.

A simplified model with lumped impedances has been created, shown in figure 4.14, to identify the cause of the problem. Capacitive and resistive values are estimated from the layout geometries and typical parametric values specified in the AMS process parameters [45]. All resistive values are typical for routing between consecutive flip flops, except the *Rd* and *Cd* impedances which are increased to simulate the large impedance at the routing between the two halves of the layout.

In accordance with previous assumptions, the results of the simulation in figure 4.15 clearly show that flip flop *D4* always skips a one following a zero as the output stays low for an extra clock cycle compared to the previous outputs. The cause of the malfunction is evident in the simulations of figure 4.16 and 4.17 showing the response of the internal clock signals of flip flop *D3* and *D4* together with the input and internal node *A* of flip flop *D4*. The first plot show a correct data transition from zero to one, while the second show an erroneous data transition from one to zero. In both simulations, node *A* (see figure 4.13) should keep its value after the clock tick because when *cn* is high, node *A* should have the inverse value of the flip flop input. But because the data input switches simultaneously with the clock (or preferably a bit later to satisfy hold demands), node *A* should keep the value when *cn* goes low and the inverter is shut off.

In the first plot, node *A* barely manages to stay high after the clock tick, and the flip flop works correctly. But in the second plot the same node rises while it should have stayed low for another half clock cycle. The only difference from the first plot is that the input goes from high to low. This shows that the error is asymmetric causing zeroes

to jump a clock cycle more easy than a one, effectively eating ones preceding zeroes, which is consistent with earlier measured results.

Problem solution

The simulations indicate that the problem is marginal because it only occurs in some of the cells, and it only affects ones. To solve the problem it is probably sufficient to reverse the direction of the clock feed making the clock delay in favor of the data propagation as the last bits are switched first. But to make sure a reasonable high scan frequency can be used, it might be wise to add clock buffers in addition to reversing the clock feed.

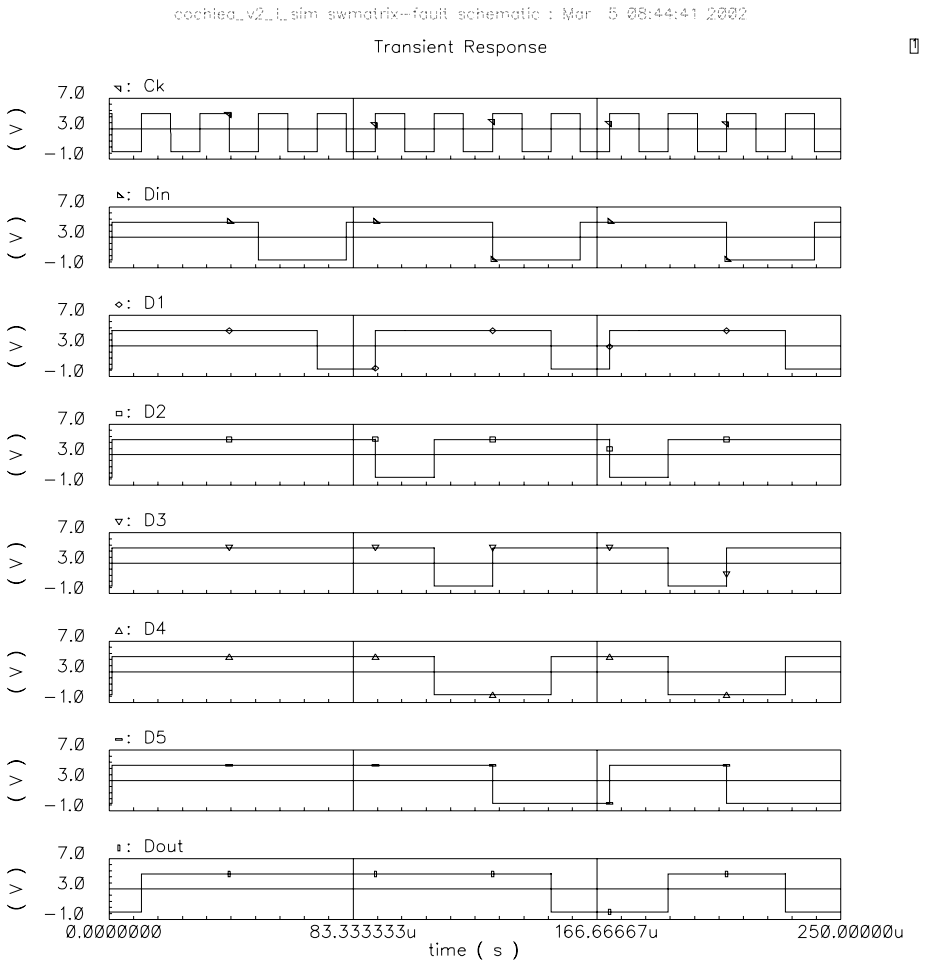


Figure 4.15: Simulation results showing propagation error of the data signal. Note that one following a zero is converted to zero at output *D4*.

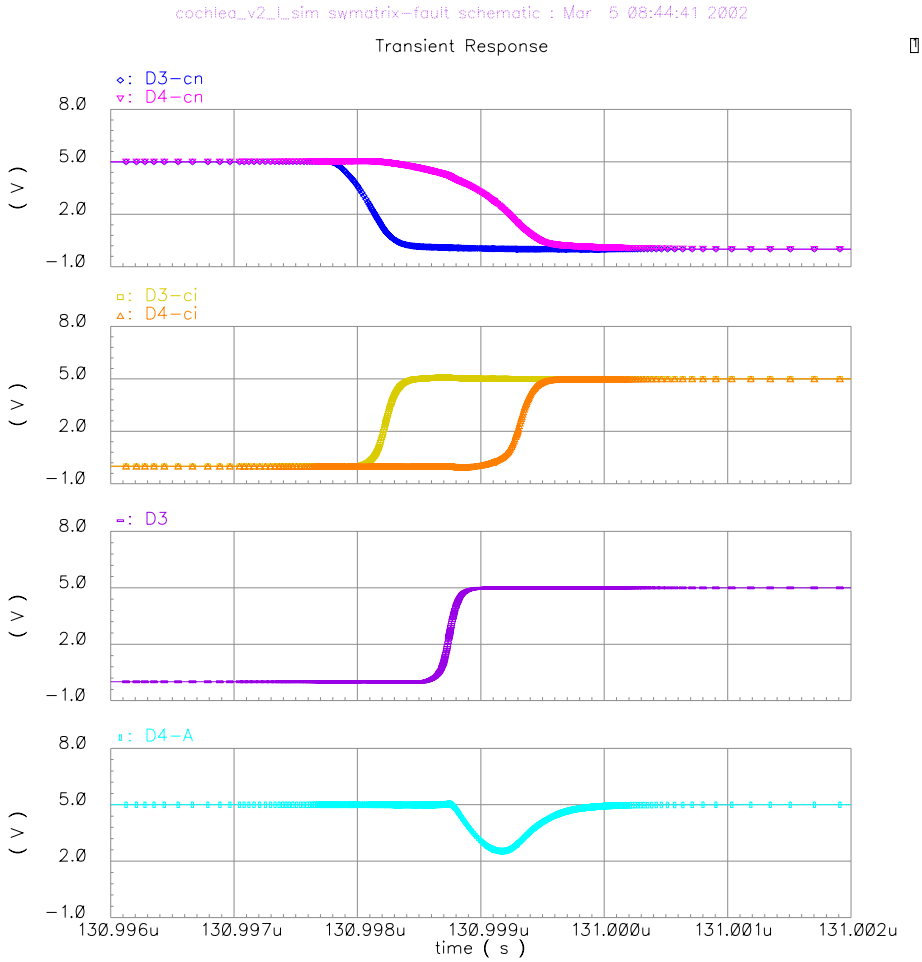


Figure 4.16: Internals of output *D3* and *D4* during data transfer when data switches from zero to one. Note that output node *A* of flip flop *D4* correctly stays high.

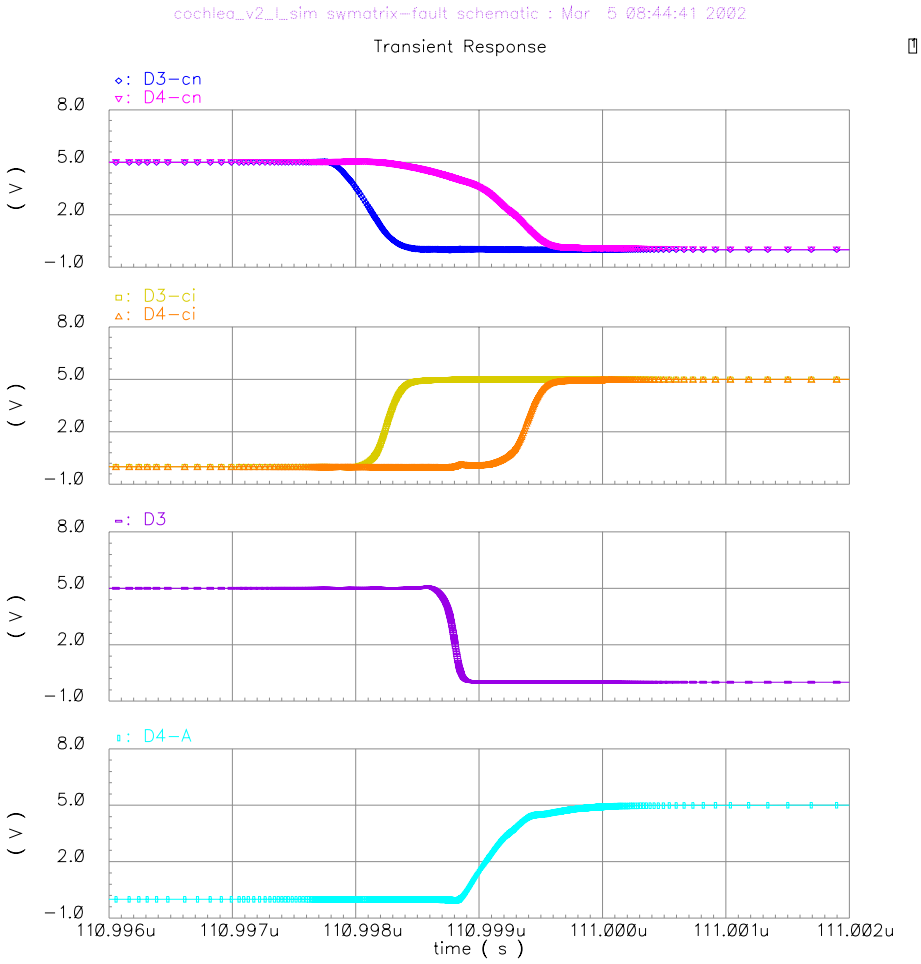


Figure 4.17: Internals of *D3* and *D4* during data transfer when data switches from one to zero. In the bottom figure the input inverter of *D4* reads the new value of *D3* a half clock cycle to early due to the delay of the internal clock signal *ci*.

4.4 System simulation and measurements

A few system simulations and measurements have been performed on the complete mixer with neurons encoding the signal. The results are described in article [30] in part II, and summarized in this section.

All simulations are executed with a MATLAB script modeling a neuron bank with linear inhibitory feedback and OR-ing of the output pulses. To simulate an asynchronous system, the maximum spike rate was held low compared to the finite system time resolution. The code is included in appendix B.

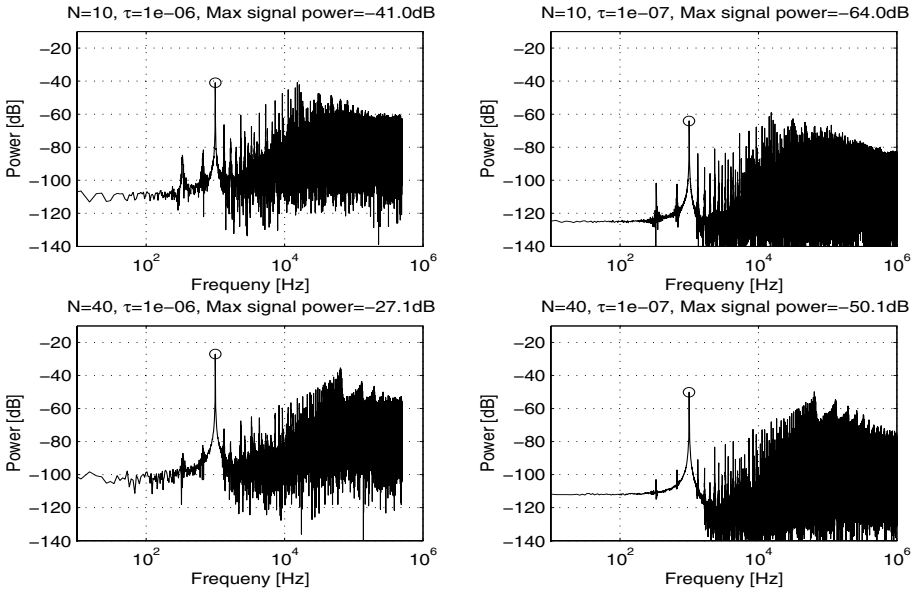


Figure 4.18: Normalized MATLAB simulations showing noise shaping for 10 neuron network and 40 neuron network with time resolution of 1μs and 100ns.

	$f_s = 1$**MHz**	$f_s = 10$**MHz**
N=2	SNR = 27dB	SNR = 31dB
N=10	SNR = 42dB	SNR = 46dB
N=20	SNR = 49dB	SNR = 53dB
N=40	SNR = 56dB	SNR = 60dB

Table 4.4: Summarized results from MATLAB simulation. SNR is computed difference in noise power in the frequency range [100Hz, 10kHz] and maximum signal power (1kHz).

As expected, the signal power increase with the number of neurons. According to the simulation results shown in figure 4.18 and table 4.4 a doubling of the number of neurons increase the signal to noise ratio and signal power by 7dB. Unfortunately, accurate measurements was difficult due to the problems with the voltage to current converter

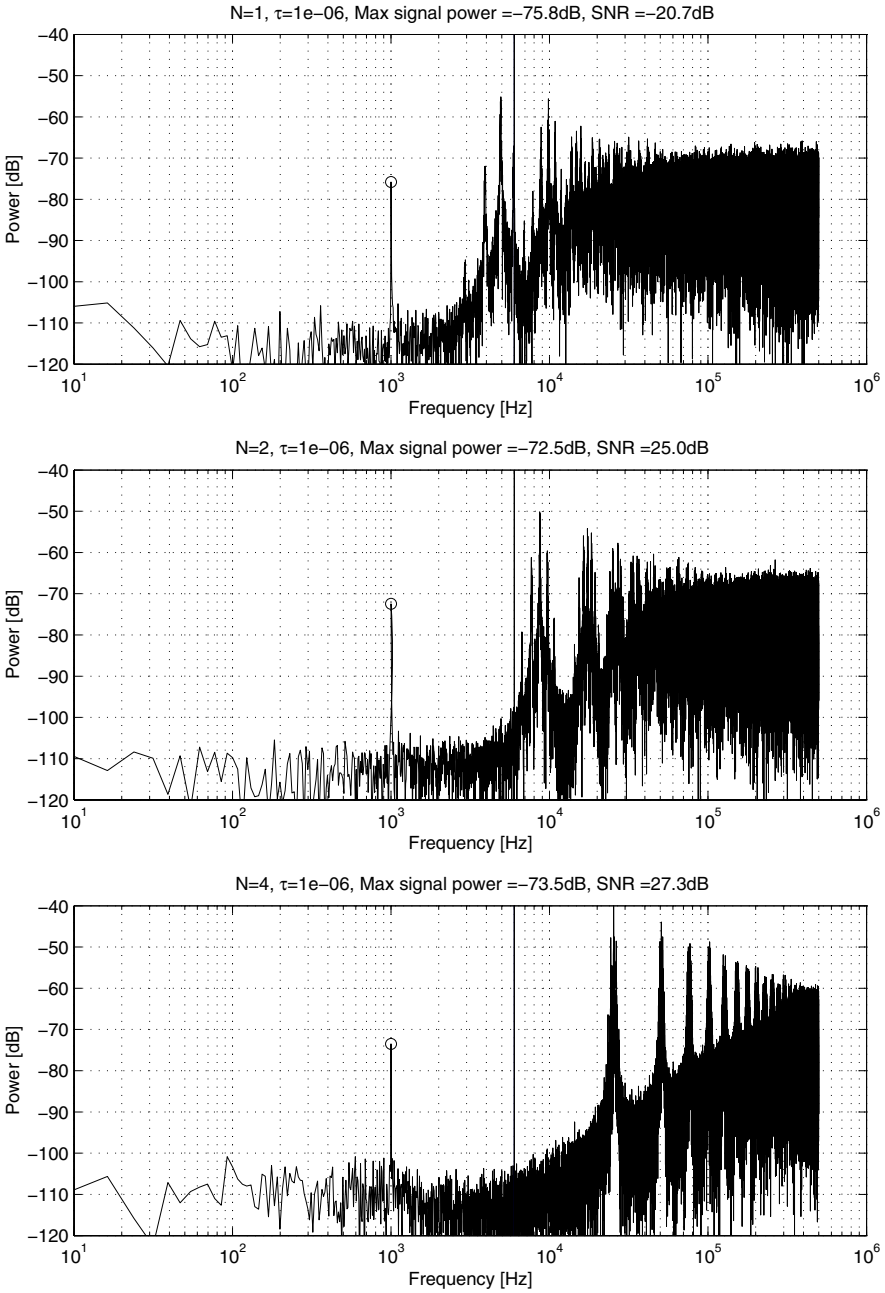


Figure 4.19: Measurements showing results for one, two and four neurons with a sampling frequency of 1MHz. The indicated SNR ratio is computed for the entire signal band below 6kHz, as indicated by the solid horizontal lines.

described in chapter 3, thus this number has not been verified by measurements. Even so in figure 4.19 measured results are shown, showing the expected SNR trend although the power increase is much smaller. The main reason is probably that the resulting spiking rate of the neurons varies greatly due to different current scaling.

Even though a deficiency of the test chip limit the ability to verify the simulated result, it is likely that the neuron-network itself is working. From the figure it is evident that the shaped noise is moved up in frequency when the number of neurons is increased. The earlier assumed variation of the neuron input current scaling (section 3.1) should only cause a slight change in signal power if the primary neuron has a relatively large input current, but still the noise shaping is improved since more neurons are contributing to the computation. Without the global inhibitory feedback no improvement in the noise shaping will be visible when adding the output of multiple neurons.

4.5 Conclusion

Principles from neuronal systems have inspired the proposed scalable mixer. The key for sufficient SNR is a novel inhibitory feedback network making it possible to distribute the noise shaping over a neuron population to increase the signal to noise ratio far beyond the achievable figure of a single neuron. From simulations it is found that a doubling in the number of neurons leads to a 7dB SNR increase, making it possible to achieve a good SNR with a relatively low number of neurons. Unfortunately the possibility of system measurements was very limited by the mismatch of the voltage to current converters described in chapter 3 reducing the possibility of verifying the implemented system. Still, MATLAB simulations show very promising results which are partially supported by the limited measurements performed.

With the novel digital switch matrix, the outputs of the neuron bank are mixed to create patient dependent sound profiles. The simple serial shift register permits sufficiently fast and simple reprogramming of the profile. A problem with the serial programming interface was identified, and a very simple solution was proposed.

The main benefits of the system consisting of the neuron-bank and mixer are the simple programming interface, small size, low power consumption, as well as very good scalability. The complete system is created from a small library of sub-cells, reducing the overall design time since the performance of each cell is independent of the size of the complete system.

Still, more measurements should be performed to reveal all properties of the system. Of particular interest are measurements showing the system performance for different weighting of the feedback, and the importance of properties such as delay on the axonal feedback.

Chapter 5

Conclusion

5.1 Designing a complete cochlea implant sound processor

The previous chapters propose an efficient implementation of the core of a cochlear implant sound processor based upon an accurate model of the human cochlea. Mainly analog circuit design is covered, but there are numerous references and parallels to medical disciplines in attempt to unite the better of the two disciplines. In short, the cochlea core consists of an analog sound processing unit which divides the sound into a large number of frequency bands. These bands are in turn processed in the spike domain to compress the number of frequency outputs to fit the number of accessible taps in the patient's electrode array. The main benefits of this strategy are compact and efficient analog implementation of the computation intensive signal processing, still maintaining digital configuration for patient adaptation of the device.

In the following sections the main contributions of this thesis is summarized.

5.1.1 Neuromorphic signal processing

In section 4.1 the concept of neuromorphic signal processing is introduced. The question is if signal processing can be inspired by how the human ear and brain appears to process sound. Perhaps the most important feature is encoding of the sound into a digital signal, much like what is done in most electronics today which tends to convert all analog signals to the digital domain when possible. But unlike most conventional electronics, the neurons generate an asynchronous pulse stream which is independent of a global system clock. The benefits are that sensitivity to amplitude noise is practically removed, and signal mixing is simplified compared to mixing of pure analogue signals.

5.1.2 Interpretation of neuromorphic coding

Several articles suggest coupling between neuron axonal encoding and $\Sigma\Delta$ encoding. This thesis uses a simple block-diagram to show that neurons might be understood using conventional $\Sigma\Delta$ converter theories, see section 4.1.1. In addition, the differences of neurons and $\Sigma\Delta$ converters are discussed. In particular the neurons are inherently stable by their architecture, in contrast to $\Sigma\Delta$ converters which must be carefully designed to

assure stability. This feature is very beneficial when making a large scale robust system. It could even be exploited to easily adjust the conversion gain without jeopardizing converter stability.

5.1.3 Patient adaptation

It is a requirement that the device must be adaptable to the individual patient by permitting adjustment of both the signal power and frequency of the individual channels. When using traditional digital signal processors, these features are simple to implement. With analog electronics, programmability is usually more difficult to implement since it often requires storage of analog voltages. But with the implementation of the switch matrix proposed in section 4.1.2, a completely digital configuration is possible. Thus the implementation utilizes the benefits of an analog implementation still providing a simple digital programming interface without any need for analog storage or additional digital to analog conversion circuitry. Note that this implementation is relying heavily on the novel inhibitory feedback network described in section 4.1 to open for direct mixing of the neuron axonal pulses, and to maintain an acceptable SNR throughout the system.

5.1.4 Improved DC offset correction

To improve scalability of the cochlear model, an improved DC-offset correction circuit was implemented. In particular the necessary filter order was elaborated upon, and measurements show that the circuit performs as desired, ensuring distortion free operation of a 100 output cochlear cascade. See section 2.6 and article Improved DC-Offset-correction in analog cascade-filters.

5.1.5 Power and space consumption

Perhaps the major challenge of a cochlea implant signal processor design is power consumption. These devices are worn by the patient and must under normal operating conditions be battery powered. Since the trend is to miniaturize the devices to make them wearable behind the ear (BTE), in the cochlear duct, or even completely implanted, it is of major importance to reduce power-consumption to an absolute minimum to maintain an acceptable operation cycle before battery re-charge or replacement.

Another trend is to increase the number of electrodes of the implanted device to improve the precision of spatial stimulation. To utilize the added electrodes, the processor core complexity must normally as well be increased together with both power and space consumption. The proposed implementation does already have a very large number of internal channels, but the number might be increased if increased frequency resolution is desired. To support a large number of electrodes, only an increase in the number of outputs from the switch matrix is necessary, only modestly affecting the power and space consumption.

5.2 Remaining implementation issues

The proposed device is not a complete cochlear implant, but if the earlier discussed fault issues of the switch matrix are corrected, it will be capable to do the signal processing for medical trials. A discussion of necessary additions for different configurations is included in the following sections.

5.2.1 Telemetry

Communication with the electrodes depends how the device is used. In a test fixture the chip will typically communicate with some circuitry provided by the lab before the signals reach the implanted device. Necessary adaptation of the signals from the test chip is test fixture dependent, and can generally be implemented by the FPGA card used in the chip measurements for sampling of the digital pulses.

If on the other hand the device is to be used in a commercial implant, the level of necessary communication interface electronics depends on the implant type. Externally worn signal processing units convey the signal through an electromagnetic coupling to the internal device. This strategy requires both signal modulation for the inductive transmission, and implementation of a higher level communication protocol.

A more compact solution is to implant even the sound processing unit. This would typically be done by placing a microphone behind the patient's eardrum and supplying the device with power from a rechargeable battery. This battery can be charged when the patient is sleeping through a simple inductive coupling. One of the major challenges with this implementation is the power consumption and the need for periodical replacement of the battery. It is not sufficient to reduce the power consumption of the signal processing core, even an improvement in electrode connectivity is necessary to reduce the relatively large currents necessary for satisfactory nerve excitation with common electrode arrays. Fortunately there is ongoing research to refine the electrode array.

Except for the practical and cosmetic benefits of a fully implanted device, there is even a benefit of simplified electronics since the telemetry interface is superfluous. This save power during normal operation, reduce the complexity of the device and removes the bandwidth limitations of the interface.

5.2.2 Electrode access

For evaluation direct electrode access is desirable because it permits full control of pulse amplitudes and in particular the pulse onset. It provides the opportunity to study the ability to convey more information to the patient if simultaneous or overlapping pulses to different electrodes are allowed. Yet another useful study is to determine the effect of spatial inhibitory feedback in the neuron-bank. This permits control of pulse-collisions, and might be adjusted to ensure a minimum electrode distance between two simultaneous events, or even generate a collision free pulse code.

To permit a higher degree of freedom for adjustment of the inhibitory network, it is possible to implement an additional arbitration circuit on the output-taps of the switch-matrix, used solely to control pulse-collisions. One extreme is to queue all events, which introduce substantial relative delay. The other extreme is to delete all colliding pulses but to maintain precise relative pulse spacing. See [3, 17, 21, 31, 34, 35] for examples

of implementations utilizing these principles. Both these methods might work well with large redundancy in the signal encoding, but this is not the case for cochlea implants.

A better alternative is to trade pulse loss with timing distortion. Assuming both the presence of a pulse and its relative timing to adjacent pulses are information carrying parameters, there will exist a threshold where the pulse delay introduce more error than a complete loss of the pulse. A design with the desired capabilities has successfully been demonstrated on an ORBIT 1.2μ process, as described in e.g. [25, 24, 23], and partly in AMS 0.8μ [32]. All the schematics are readily available, and only some adaptation of a few component dimensions and re-simulation of a two critical sub-circuits should be necessary to adapt the design to the current process.

5.2.3 Digital control

A digital control circuit is necessary to be able to control the switch matrix. This matrix is configured through a serial interface on the current test chip, but the configuration bit stream must be downloaded from an external source. To complete the configuration circuitry, a reprogrammable memory, some control logic and a clock generator should be included. The memory is used for storage of one or several patient specific configurations, while the logic and clock generator controls the programming cycle. With a suite of patient specific programs, the user is allowed to switch between different settings for different listening conditions.

5.3 Concluding remarks

An obvious interest area that should be pursued is medical trials using the sound processor described in this thesis. With identified corrections and possible external interface adaptations of the existing implementation, important features of the proposed sound processing strategy can be tested. And with addition of frequency dependent gain adaptation, it should be easier for the user to discern desired sound features in a noisy environment.

There are two likely short term benefits of such trials. First it can be revealed if the model itself has any benefits over existing models in terms of sound quality. If so, this knowledge should be used in future development of processing algorithms implemented on more traditional cochlear implants. Second, if the chip performs as expected, this will solve power consumption issues and potentially be a base for development of a fully implantable device. Note that the close match to a biological cochlea of the proposed model is very demanding for a real-time digital signal processor-based system due to the high number of band-pass filtered channels. Even though more efficient processors are developed, there will be a benefit in implementation of a tailored processor from transistor level to maintain reasonable power consumption.

An interesting spin-off from this work is the mixer block described in chapter 4, which clearly have potential for further research. With the desirable design simplicity and inherently low power-consumption, it would be very interesting to see how such a design would perform compared to other known low-power ADC architectures. The current results are promising, and even if the architecture turn out to be unable to compete with more traditional converter architectures in terms of quality of the conversion, it

has very interesting properties such as inherent activity adaptation, robustness and self-stabilization. It is appealing to be able to implement a system interfacing with the 'real world' through a neuromorphic converter. The activity of the converter could then inherently control the activity level of the following system, avoiding complex sleep and wake-up algorithms. With intensified research activity, it is likely that these converters can find new areas of usage and open new arenas for micro-electronics.