Comparison of Neuromorphic Event-based Auditory Sensors

SCIENTIFIC SEMINAR

submitted by
Fabiola Schöller

NEUROSCIENTIFIC SYSTEM THEORY
Technische Universität München

Prof. Dr Jörg Conradt

Supervisor: M.Sc. Mohsen Firouzi
Final Submission: 06.07.2016

Abstract

Neuromorphic engineers try to mimic nature when developing electronic systems as evolution has created many concepts that greatly outperform conventional approaches. One example is the adaptable and energy-efficient biological cochlea whose performance could not be reached by any artificial auditory sensor yet. However, since the 1980s, a large number of silicon cochleae with increasing accuracy and similarity to the biological system has been presented, allowing for significant improvements of technology in many different applications, for example in speech processing.

This report reviews the historical development of neuromorphic event-based silicon cochleae, compares the performance of different basic approaches with the biological system and examines to what extend this technology is able to improve speech processing. It is shown that the main advantages seem to be lower power consumption, reduced computational cost and greater adaptability to noisy environments. We conclude by stating still existing challenges and giving an outlook on future developments.

Contents

Chapter 1

Introduction

1.1 Motivation

"Convenience" is one of the guiding principles in the development of new electronic products for the broad market. Having to use a keyboard to give commands to an electronic device is increasingly not expected from the user. Voice commands are now implemented in a lot of different applications: For smartphones, it makes handling more comfortable, allowing quick input of search requests, telephone numbers or whole emails; in cars, it enables the driver to control navigation, radio and other multimedia without having to take his hands off the wheel or his eyes off the road. Buttons are no longer necessary to operate a TV, and light can be switched off by a short call. But voice control is not restricted to consumer electronics; in industry, it allows e.g. to develop robots that can interact with humans in a very natural, intuitive way, making possible for example artificial assistance in nursing care for the growing population of elderly people.
These applications typically require reliable speech recognition in a noisy environment. For instance, the navigation system should also be able to identify commands while the car engine is running and the children are chatting in the rear. In some cases, it is also important to identify the speaker correctly, e.g. to authenticate a person that wants to do a transaction with a mobile banking app or recall medical information from a telemedical system.

Current state-of-the-art technologies do propose solutions to this complex task; hidden Markov models (HMMs) that capture temporal dependencies in the acoustic pattern "have been immensely important in the development of large-scale speech processing applications and in particular speech recognition" [AL11] as they are very flexible towards the length of the input sequence, but their performance suffers in noisy environments. HMMs model acoustic and temporal variability in speech via statistical probability distributions, however "even the best systems are vulnerable to spontaneous speaking styles [and] non-native or highly accented speech." [GY08]. Another current approach are the Support Vector Machines (SVMs) that use learn-

ing algorithms developed in the machine learning sector. In spite of being a powerful technique for many pattern recognition problems, they are restricted to a fixed-dimension input which makes it difficult to use them for speech recognition and they "cannot model the temporal structure of speech effectively." [LW05].
Humans themselves perform much better in the task of analysing speech. In several hundred million years, evolution has developed a system that can work with a huge range of sound intensities by amplifying or compressing if necessary, and analyse frequencies, volume, direction and distance of incoming sounds very energy-efficiently and in many different listening environments, also in noise. This leads to a speech recognition ability that surpasses every technical system. However, biology achieves this performance in a completely different way: "In machines, acoustic signals are typically chopped into discrete frames from which a set of extracted features are subsequently classified by machine learning techniques. In biology, the input is first processed in continuous time by the cochlea, and the resulting stimulus-driven asynchronous spikes from the cochlea are distributed to various auditory brain areas." [AL11]. Therefore, neuromorphic engineers try to model in silicon the robust and effective cochlea and subsequently more parts of the auditory pathway by mimicking as close as possible. However, this is a big challenge as not even the biological cochlea has yet been completely understood. [AL11, LDL12, USH06]

In the following, we will first introduce the biological cochlea and its functionality. Subsequently, the historical development of silicon cochleae will be reviewed by presenting and comparing selected examples that show the progress in architecture and performance. Finally, we will examine, in the first instance, to what extend the latest sensors meet the expectations of modern speech recognition and speaker identification, and furthermore shortly address the logically following question of whether neuromorphic cochleae could provide an opportunity to replace a damaged biological hearing organ with an implantable chip system.

1.2 The Biological Cochlea

In this section, we will present the structure and especially the functionality of the biological cochlea in order to make the approaches plausible that were used by the engineers of silicon cochleae.

The main task of the biological cochlea is to convert the pressure signal of the incoming acoustic sound into neural signals that can be evaluated by the brain. Figure 1.1 provides an overview of the whole transduction system that will now be further explained.
The cochlea is a fluid-filled duct in the inner ear that spirals from the base to the apex. In the cross-section at the lower right corner of figure 1.1, three chambers are visible; the scala media sits in the middle and is isolated from the scala vestibuli

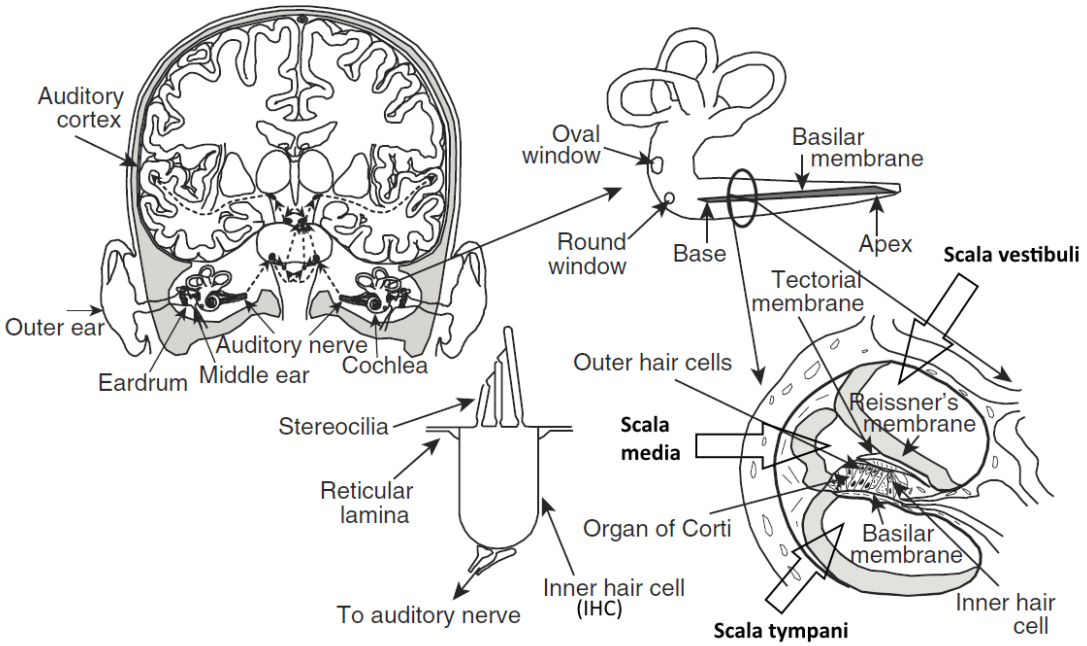

Figure 1.1: Overview of the auditory pathway (from [LD+15], edited)

(above) by the Reissner's membrane and from the scala tympani (below) by the Basilar membrane (BM). The scala vestibuli and the scala tympani are connected at the apex of the cochlea, while the scala media is separated from them and contains a fluid with different ionic composition than the other two chambers.
The physical properties width and elasticity of the BM are not constant in the whole cochlea: At the base, the membrane is narrow and stiff, whereas at the apex, it is wide and flexible. This leads to a greater movement in response to high frequency stimuli at the base and to low frequency stimuli at the apex, and therefore, if the input consists of a sound with several frequencies at the same time, to a spatial separation of the higher and lower frequency components along the cochlear axis. The frequency that causes the greatest oscillation at a particular place along the BM is called characteristic frequency of that place, so the characteristic frequency decreases from base to apex. In numbers, it declines from about 20 kHz to 20 Hz.
On top of the basilar membrane, there is the organ of Corti that contains the inner and outer hair cells (IHCs and OHCs). An IHC is depicted in 1.1 in the bottom center position. Both kinds of cells have a hair-like structure at their tips, the stereocilia, and are connected to the brain via nerves to report deflections caused by a sound pressure wave.
The vibrations of an incoming pressure signal are transferred from the eardrum to the oval window of the cochlea by three small bones in the middle ear. As the fluid within the cochlea is considered incompressible, both Reissner's membrane and the

BM are deflected and the round window compensates the pressure by performing opposite movements. If a pure tone with a certain frequency is presented, "the displacement pattern [of the BM] builds up at the base of the cochlea until it reaches a maximum amplitude at the place along the BM where the pure tone frequency is [equal to the local] characteristic [frequency]. It then falls off quickly as it travels towards the apex of the cochlea." [Ham08]. This pattern is called travelling wave.
The displacement of the BM that depends on the characteristic frequency at the particular place and the sound's intensity in turn cause a shearing force within the organ of Corti and therefore a deflection of the stereocilia of the hair cells. The IHCs encode the characteristics of the sound via action potentials: Intensities in the range of 120 dB SPL are encoded with spike rates, temporal information is transferred by phase-locking the individual cycles of the sound wave if its frequency is below about 4 kHz, and the frequency of the tone can be determined from the characteristic frequency of the position of the IHCs that fire most; in case of a multi-frequency sound, superposition applies. The action potentials are then carried to the brain by the auditory nerve.
The main task of the OHCs, according to the latest scientific findings, seems not to be a transfer of the sound signal into neural code, but a regulation of the BM's movement. At low input intensities, they alternately reduce and increase their length in response to the back and forward movement of their stereocilia, therefore adding in-phase energy. At high input intensities, they increase damping. This leads to a couple of phenomena: First, the response of the cochlea can have more energy than the sound signal itself, which can lead up to the emission of sounds by the inner ear that are called otoacoustic emissions. Second, "it was observed that the gain at the characteristic frequency increases as the input decreases and vice versa" [Ham08], reaching saturation for high input levels. This is shown in the right diagram of Fig. 1.2. However, "at frequencies lower than the characteristic frequency the relationship between input and output was observed to be linear." [Ham08]. This leads to a sharpening of the travelling wave around the characteristic frequency which increases selectivity and can be seen in the left diagram of Fig. 1.2. [Ham08, Shi04]

Now, as the main parts and key features of the biological cochlea have been explained, we will focus on how the achievements of biology can be transferred into technology in the following.

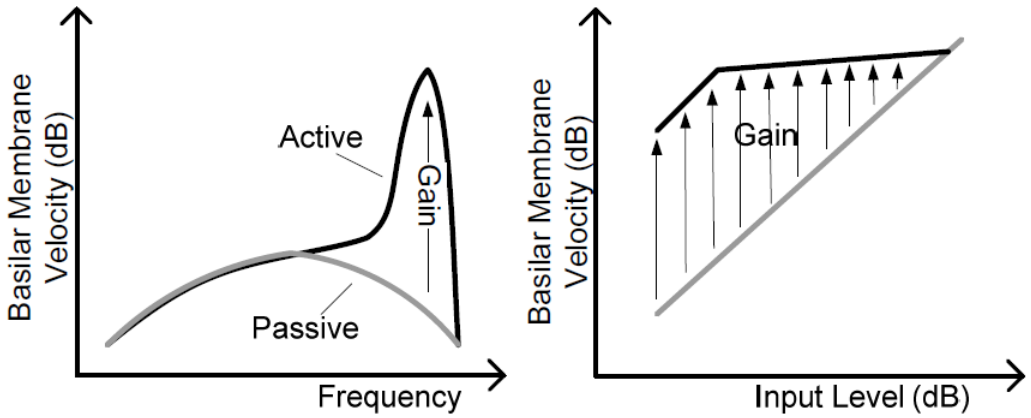

Figure 1.2: The effects of the OHCs on the basilar membrane velocity; left: Increased gain at the characteristic frequency leads to higher selectivity, right: amplification of the BM velocity at low input levels, damping at high input levels (from [Ham08])

Chapter 2

Hierarchical Development

The basic modules for a silicon cochlea are transistors operating in the subthreshold regime, as here the current depends exponentially on the voltage, similar to the exponential dependence of the current through voltage-sensitive ionic channels on potential differences across a neuronal membrane. The introduction of low-cost analog very large scale integration (VLSI) in the 1980s was a starting shot for the development of neuromorphic silicon cochleae. [LD10, WB11]

All architectures involve a number of filters or resonators that, out of technical reasons, model a discretized BM as every filter can only simulate one characteristic frequency at the same time. These frequencies decrease exponentially from base to apex, similar to the biological original. The change in stiffness and width of the BM with longitudinal position is generally modelled by adjusting the parameters of the cochlear filter elements.
However, the coupling between the elements is realized differently which leads to a classification in 1D if only the longitudinal wave propagation along the BM is modelled, or 2D if also the vertical wave, caused by the propagation within the fluid, is taken into account; but combinations of both extreme cases are existing as well. Furthermore, in active cochleae, gain and frequency selectivity vary depending on the input intensity, whereas in passive ones not, which neglects the effect of the OHCs, but reduces complexity.

In the following section, we will give an overview of some important historical milestones in the development of silicon cochleae, from the first cascaded 1D models, active 2D derivatives, mixed-signal chips with active bidirectional coupling, including silicon neurons, up to architectures including binaural information.

2.1 Historical Development of Silicon Cochleae

1D cascade models The first silicon cochlea, built in CMOS VLSI technology, was put forward by Lyon and Mead in the year 1988 ([LM88]). It consists of a

cascade of 480 second-order filter sections that models an array of BM segments, each with the physical length Δx. The characteristic of the filters is approximately between low-pass and band-pass characteristic and they differ only in their quality factor Q and their time constant τ (the inverse of the characteristic frequency). τ is increasing exponentially from the beginning of the cascade to the end. As the input signal travels through the cascade, its high-frequency components are selectively filtered out, such that at the end of the cascade, only the lower frequency components remain. Each filter therefore shows a response displaying a steep cut-off of all frequencies higher than its characteristic frequency, "due to the successive removal of the high-frequency components by the filters before it" [LD$^+$15]. In Fig. 2.1, the diagram on the right depicts this effect (solid line) in comparison to a stand-alone second-order filter section (dashed line). On the left, a schematic of the 1D cascade cochlea model is presented. [LD$^+$15]

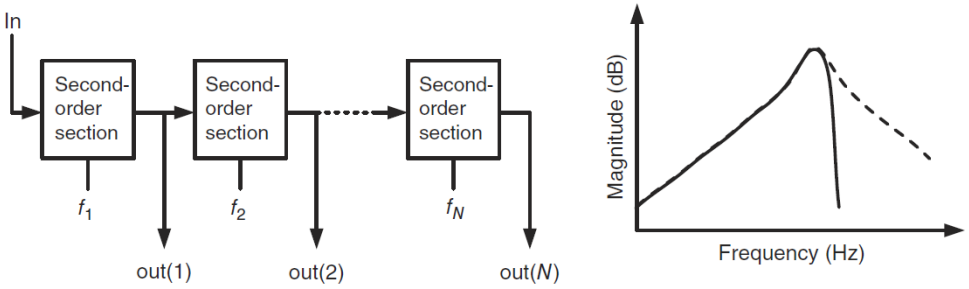

Figure 2.1: The 1D cascade cochlea model (left) with N filter sections, each having their characteristic frequency at f_i, and a comparison of the output of one of these filter sections (right, solid line) with the output of a stand-alone second-order filter section (right, dashed line) (from [LD$^+$15])

Since its publication, this first model of Lyon and Mead has been picked up and further improved many times.
In 1992, Watts, Kerns et. al [WK$^+$92] published a version with an increased linear range due to better exponential spacing of the characteristic frequencies; also large signal stability and matching of the quality factor Q between the sections could be improved.
Van Schalk et. al [vSF$^+$96] succeeded 1995 in further uniforming the frequency responses at different output taps by including compatible lateral bipolar transistors.

However, all 1D cascade types have some major drawbacks: One issue is that a failing section makes the rest of the cascade unusable, as each filter output drives the input of the next filter. Then, within the cascade, noise immensely accumulates, [WB11] exemplary named a factor of 100; the number of stages is limited due to a an increase of delay with the number of filter sections and also dynamic range and large signal stability are problematic. [Ham08]

But "silicon cochleas that consist of 1D parallel sections or a 2D structure circumvent some of the major drawbacks of the cascaded filter sections." [LD+15]. Therefore, we will now focus on 2D models only.

2D models The basic difference between the 2D approach and the 1D cascade model is that the cascade is replaced by a resistive network which is a model of the cochlea fluid and can be seen schematically in Fig. 2.2. The dependency of each filter stage on the output of the filter in front of it is resolved, which leads to a faster and less error-prone behaviour.

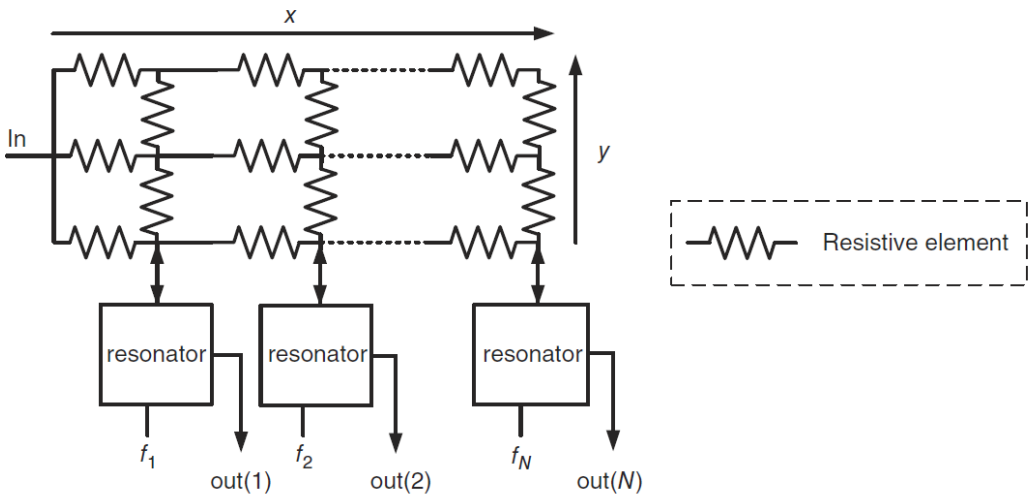

Figure 2.2: The schematic structure of the 2D silicon cochlea with a resistive network modelling the cochlear fluid (from [LD+15])

One commonly used model, first published by Fragnière [Fra98] in 1998, was translated into a software implementation and simulated with an input between 100 Hz and 10 kHz by Shiraishi [Shi04]. The response shows the sharp cut-off after the characteristic frequency, as well as a greater quality factor Q for higher frequencies and is therefore close to the response of a passive biological cochlea, i.e. with inhibited OHCs.

Active Models Until now, one major lack of the silicon cochleae presented was their passiveness which means they neglect the effect of the OHCs in the biological cochlea and do not adapt to the intensity of the input signal to increase their signal-to-noise-ratio. One approach of including the active nonlinear behaviour of the OHCs is the automatic quality factor control (AQC) which was first proposed in Lyon [Lyo90]. Hamilton's model which is using AQC [Ham08] shows for decreasing input intensity, an increase in gain and selectivity, an upward shift in the characteristic frequency as well as nonlinear behaviour such as distortion product otoacoustic

emissions in response to combinational tones; these effects all appear in the biological cochlea, too. However, in comparison, input dynamic range and gain are still considerably limited in the silicon model. [LD+15]

An active bidirectional coupling model In order to further improve the performance of silicon cochleae, developers try to better imitate biology by modelling more details of the natural architecture. In the following, we will present an approach of mimicking the cochlea's micromechanics, i.e. the anatomy of the organ of Corti, and including silicon neurons in more detail.
In 2009, Wen and Boahen presented a mixed-signal VLSI cochlea chip with active bidirectional coupling (ABC) [WB09]. The reason for using a mixed-signal architecture was to combine the advantages of space-saving analog design with digital VLSI that makes interfacing with higher-level processing easier. The disadvantages of analog elements like mismatch and noise were conjectured to be compensated by using ABC, "given that the biological cochlea itself is built out of imprecise components" [WB11].
The basis for ABC is the observation that in biology, the OHC forces are also transmitted both forward to an adjacent downstream BM segment and backward to an adjacent upstream BM segment. Additionally, the model integrates nonlinearity that is assumed to be originating from a saturation of OHC forces, because physiological measurements show that both the OHC's receptor potential and the change in an OHC's cell body length saturate with the acoustic pressure. [WB11]
The design was implemented by first creating a passive model in which the cochlear fluid was represented by MOS transistors and the BM by second-order systems with two low-pass filters each. Then, ABC with saturation was included to create a nonlinear active system. Therefore, two saturated and scaled currents from both adjacent neighbours were added to the input of the second low-pass filter in each BM segment. Fig. 2.3 depicts the described BM circuit for one BM element: The two boxes left and right are the low-pass filters, and the communication with the neighbours via current is shown.
The analog output signals I_{mem}^{+} and I_{mem}^{-} are then transformed into digital pulse streams, representing the spike trains of the auditory nerve fibers, by silicon neurons. They consist of IHC circuits that create half-wave rectified, low-pass filtered currents which drive 6 spiral ganglion cell (SGC) circuits each. The silicon SGCs encode the temporal features of the sound stimulus, i.e. at stimulus onset, they fire at a higher rate, and they phase-lock an incoming sinusoid. Finally, the output of each SGC circuit is transmitted off-chip by an address event encoder based upon the address event representation (AER) interface protocol which we will describe in the following. [WB11]

In the AER protocol, every neuron gets a unique address so that spikes can be represented with a spike address containing information about the time and location of the event. This spike is carried off-chip by a shared address bus and can be e.g.

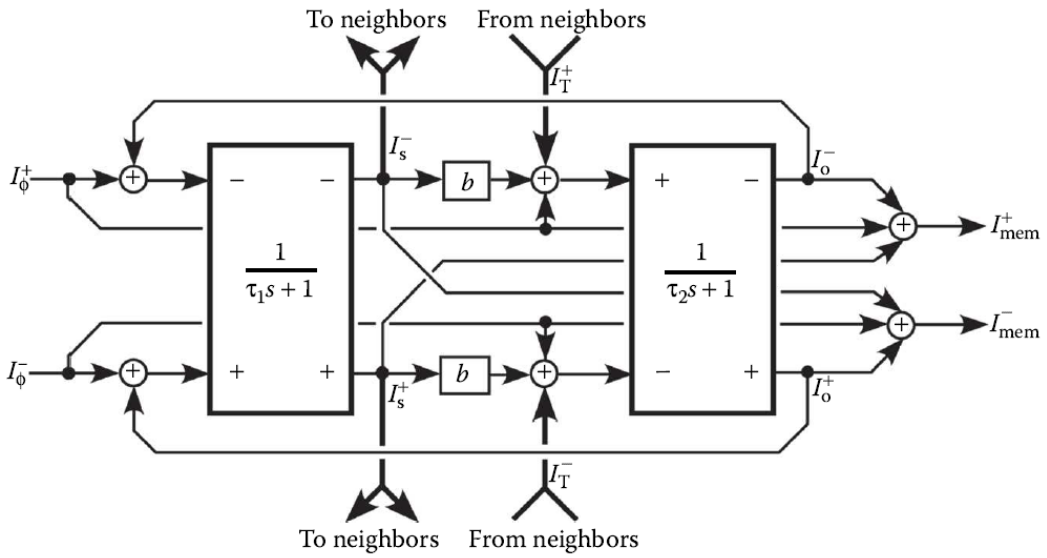

Figure 2.3: Circuit for one BM element, containing the two low-pass filters and the connection to the adjacent neighbours via current exchange (from [LD+15])

transmitted to another neuron with the help of a routing table stored on a digital memory chip. This means on the one hand that the connectivity pattern between the neurons can be changed by software and on the other hand that "this method allows multiple transmitters to communicate with multiple receivers in a pseudo-parallel fashion." [CLVS07].
AER has two main advantages compared to other systems: Firstly, as every cochlea channel can transmit events autonomously, there is no redundancy in information, which reduces "power dissipation both for communication and for subsequent processing." [LD10]. Secondly, latencies are very short as there is no regular external clock pulse that controls the time a transmission can happen, so post-processing can promptly be initiated. So with AER, the timing between events is statistically preserved in a way that models biological synaptic transmission very well. [LD10, CLVS07]

Coming back to the model of Wen and Boahen [WB11], according to the AER protocol the neuron, which is represented by the spiral ganglion cell SGC, creates a request by pulling the request line low.
Therefore, it is required that the acknowledge signal is low. An on-chip arbiter, here the address event encoder, sends the spike address off-chip and an acknowledge back to the neuron, where the acknowledge line is pulled up and the neuron resets. On the receiver chip, the address is decoded and the target neuron receives the spike. [WB11, CLVS07]

Wen and Boahen [WB11] fabricated this implementation with 360 BM-segment circuits, hence covering a frequency range of 200 Hz – 20 kHz and an input dynamic range of 52 dB, and tested on pure and complex tones.
In the first instance, the performance of single BM segments was evaluated by measuring the current outputs. For different frequencies, the responses shows some irregularities in response shape and peak height due to transistor mismatch, but in general, biological characteristics are reproduced.
Next, measurements of how the travelling wave amplitude builds up revealed that ABC indeed works in a distributed way. However, the responses largely vary between segments, again because of transistor mismatch.
For increasing input amplitude from 0 dB to 48 dB, "BM responses show compressive growth, first at the CF [characteristic frequency] and then at nearby frequencies" [WB11], which leads to a more broadly tuning without a significant phase change, like in biology. But below and above the characteristic frequency, compression is bigger and not symmetrical around the peak; additionally, at intensities above 48 dB that could not be tested, the response of a biological cochlea would become linear again.
By varying the coupling's saturation level, it was demonstrated that "ABC captures the role of OHC electromotility–at least qualitatively." [WB11]. The mechanism successfully prevents accumulation of noise, but its gain increase is "limited to 18 dB by mismatch-induced traveling-wave reflections." [WB11].
Finally, evaluating the spike trains of the SGCs in response to a chirp-click sequence revealed that the chip succeeds in encoding the sound's frequency via place code, intensity via rate code and timing via real-time code. Again, transistor mismatch is an issue as it makes some SGCs too excitable.

AER EAR So far we presented only single cochlea architectures. However, in 2007, Chan, Liu and van Schaik [CLVS07] matched two silicon cochleae to AER EAR to still further approach biological hearing by evaluating binaural information as well. They applied signals with interaural time difference (ITD), i.e. they simulated a situation where the sound source is closer to one ear than to the other and the signal therefore reaches one ear earlier than the other one. As their chip uses the AER protocol (see above), it was considered possible to translate timing information into AER spikes. Applying a pure tone and, in a second experiment, white noise both directly and in reverberant environment revealed that "timing information is well preserved by AER EAR and ITD can be easily extracted from the spike trains". [CLVS07].

AEREAR2 A new standard was set again by Liu, Schaik et al in 2010 [LVSMD10] as they presented a chip with local filter sharpness adjustment, mismatch robustness and easier programmability. It consists of 2 cochleae with 64 cascaded, coupled filter stages each, that contain pulse-frequency modulators (PFMs) implementing

an integrate-and-fire-model of a neuron with different thresholds to encode volume of the sound stimulus. “The PFM output addresses are transmitted asynchronously off-chip using the AER protocol. [An USB interface for control and processing] time-stamps the events to 1 µs resolution.” [LVSMD10]
The quality factors of individual channels can be adjusted and therefore filter sharpness locally enhanced by two methods: Firstly, by using local Q Digital-to-Analog Converters (QDAC), and secondly, by using a “nearest neighbour lateral inhibition scheme”. [LVSMD10]. Furthermore, the chip has binaural structure, prototype on-chip microphone preamplifiers to include global automatic gain control, and digitally controlled biasing circuits for stability against temperature and process variance which improves matching. [LVSMD10, VOR16]

This was a rough outline of the historical development of silicon cochleae from 1988 until 2010. In order to provide an overview of the differences and achieved improvements, we will now compare some models that use different architectures and shortly recapitulate the progress.

2.2 Comparison

Table 2.1 sums up main properties of four silicon cochleae presented before and of their biological counterpart.
The first type of architecture, 1D cascade, indeed offers some advantages compared to more recent models; its power consumption is very low in comparison, although this is the case because it does not mimic active nonlinear behaviour which needs energy, and it is able to cover the whole biological frequency range. However, its number of stages is limited because of the accumulation of noise and delay and therefore frequency resolution is rather poor; other drawbacks are the limited input dynamic range and large signal stability as well as the great effect a failing stage has on the rest of the cascade. But nevertheless this architecture is the simplest one and already contains important features of the signal processing within the biological cochlea.
Involving the dynamics of the cochlear fluid with a 2D architecture firstly made the model more realistic and secondly significantly improved access to the area of real-time applications, but still, the effects of the OHCs are not taken into account. The active 2D model of Hamilton eliminates this shortage by using automatic quality factor control (AQC) and therefore reproducing gain characteristics and some nonlinear features of the biological cochlea. However, this goes with a limitation of the frequency range and an enormous increase of power consumption; additionally, the possible input dynamic range is still considerably smaller than in biology.
A completely new feature was added with artificial neurons that simulate the spiking behaviour of the spiral ganglion cells. The AER protocol allows for communication between different chips in real-time, with low power-consumption compared to other

model	**1D cascade** [WK+92]	**2D Active** [Ham08]	**Active bidirectional coupling model** [WB09]	**AEREAR2** [LVSMD10]	**Biological cochlea** [SLM98]
size	2.22x2.25 mm^2	5 mm^2	3.76x2.91 mm^2	5 mm^2	length: 35 mm
number of cochlea stages	max. 51	83	360	64x2	-
input dynamic range	62.6 dB per single section, in cascade much less (no number given)	46 dB	52 dB	36 dB	120 dB
frequency range	20 Hz– 20 kHz (of unimproved model [LM88])	100 Hz– 10 kHz	200 Hz– 20 kHz	50 Hz– 50 kHz (adjustable)	20 Hz– 20 kHz
power consumption	51-stage cochlea: 11 µW, chip: 7.5 mW	net: 16.72 mW, total: 56.32 mW	analog core: 35.9 mW, total: 51.8 mW	analog core: 30 mW, digital circuits: 25 mW	estimated: 14 µW

Table 2.1: Comparison of main properties of three silicon cochlea models presented before and the biological cochlea

systems, and can at any time be adjusted by software.
In AER EAR, beside including the event-driven protocol AER, further improvements of the response could be reached by a more detailed implementation of the organ of Corti; also a significant increase of the number of cochlea stages was possible without requiring much additional space or leading to a higher power consumption, but still, input dynamic range is far away from the biological cochlea, not to mention power efficiency.
Additionally, all these models struggle with mismatch of MOS transistors operating in weak-inversion. Depending on the transistor area, threshold-voltage variations lead to current variations with, for instance in the model of Wen and Boahen, a currents' coefficient of variation of up to 25 %. Unlike the biological cochlea, the published models are not able to compensate the influence of imprecise components, and suffer from irregularities and limitations. As variance is inversely proportional to the area of the transistor, decreasing its size means increasing mismatch, but also leads to a sharper frequency tuning, lower power-consumption, broader bandwidth and smaller chips. [WB11, Kin05]
AEREAR2 however was designed for robustness against mismatch by using biasing circuits, operating in voltage-mode and locally adjusting quality factors. This makes outputs from left and right ear more precise, an important characteristic e.g. for spatial audition. The USB-interface and easier programmability offers a drastic improve in interface and flexibility. On the other hand, compared to previous models, AEREAR2 has a smaller number of channels and a decreased input dynamic range, but nevertheless, because of the named advantages, it has recently been chosen and tested for many applications in the field of auditory scene analysis.

However, whether the decades of engineering neuromorphic auditory sensors have already been sufficient for creating a superior alternative to conventional speech recognition technologies will be evaluated in the following.

Chapter 3

Discussion

Recall that main challenges for previous systems used in auditory scene analysis, especially in speech recognition and speaker identification, was their great sensitivity to noise in non-artificial environments, as well as their power consumption due to the large computational power they require and their lack of adaptability to different speaking styles. In general, they are not able to compete with human performance.

The introduction of the event-based binaural AEREAR2 with its locally adjustable quality factor and convenient programmability led to a series of tests in different, complex applications of auditory scene analysis.
In 2011, Abdollahi and Liu [AL11] published an experimental study on speaker-independent isolated digit recognition. They used the spatio-temporal spike pattern of a passive version of AEREAR2 in response to a spoken word as input for Support Vector Machines (SVMs). Therefore, while testing three different methods, snapshots of the spectro-temporal cochlea spike patterns were taken to create static image representations with fixed-dimension, so that they can be analysed by a SVM classifier. Results were compared, amongst others, to a conventional hidden Markov model (HMM)-based system. The percentage of average recognition accuracies in over 11 digit classes from TIDIGITS database reaches between 93.79 % and 95.58 % for the silicon cochlea approach, depending on the selected processing method, and 99.70 % for the HMM-based system. This shows that the conventional approach is better suited for this task which is plausible as its input does not have to be reduced into a fixed-dimension vector, while the passive silicon cochlea approach also struggles with the small input dynamic range which causes problems as the input ranges from low-energy unvoiced fricatives like /h/ to high-energy vowels like /a/. However, considering the rather simple representation of input sound via static images and the variances within the cochlear chip, this result is still remarkable. The authors plan to include “a spike-based multi-neuron learning neuromorphic chip as the back-end classifier in order to build a real-time low-power fully spike-based integrated sound recognition system.” [AL11].
While these results are promising, the system on the other hand was not tested in

noisy environments or with other disturbances, and different speaking styles were also not taken into account. Additionally, the pre-processing is, compared to biology, artificial and the active behaviour of the natural cochlea was neglected, which may lead to possibilities for further improvements.

Another study of auditory scene analysis with AEREAR2 was presented by Li, Delbruck and Liu [LDL12] in 2012. Here, the output of the active AEREAR2, more precisely the inter-spike time interval (ISI), was used for a real-time speaker identification task.

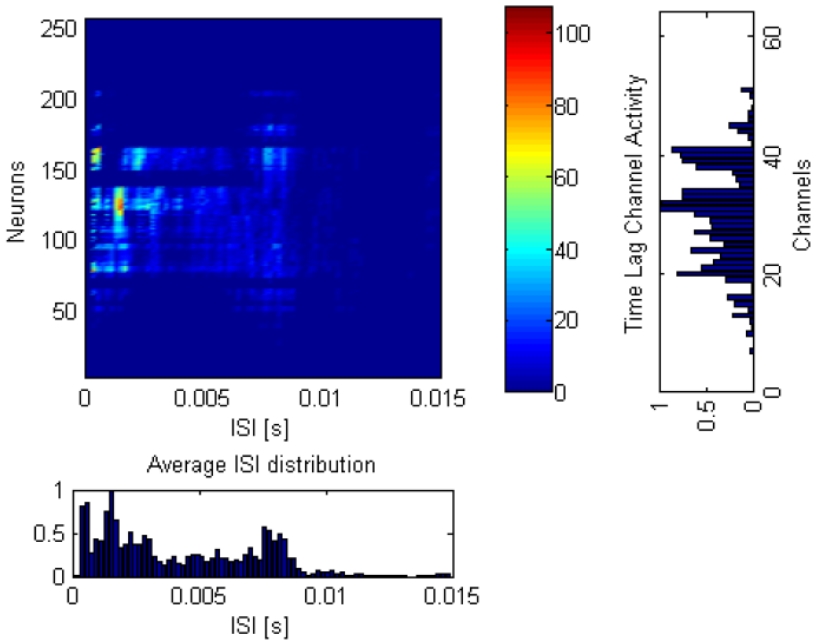

Figure 3.1: Response of the AEREAR2 to a sentence from a male speaker in the TIMIT database, from combined left and right cochlea channels. Top left: 2D feature matrix with counts for all channel neurons in an 80-bin ISI histogram. Below: 1D vector of ISI distribution across channels. Top right: 1D vector of average activity of individual channels (from [LDL12])

Figure 3.1 shows the response of the chip to a sentence from a male speaker in the TIMIT database. Both left and right cochlea channels were evaluated, although this is not necessary for the speaker identification task. Top left, the 2D feature matrix with counts for all channel neurons in an 80-bin ISI histogram is pictured; below, there is the 1D vector of ISI distribution across channels, corrected for the delay originating from the cascaded architecture of the filters, and top right the 1D vector displaying the average activity of the four neurons of each channel; computing the average leads to a noise reduction.

The two 1D vectors depicted are combined to one 1D vector and passed on to a SVM. The classifier calculates the probabilities of all learned speakers for each feature vector, and at the end of each sentence the most probable speaker. In the training phase, sentences of 0.8 to 4 seconds, spoken by 40 both male and female persons, were used. The system was then analysed by concatenating sentences from the 40 speakers and measuring the run-time decision accuracy at the end of each sentence. On average, 92 % could be reached, with a trade-off between accuracy and latency.
This study tested the chip in a more natural scenario than Abdollahi and Liu did for the digit recognition task, therefore offering a better basis for estimating performance in real-life situations, but again, environmental noise and other disruptive factors were not included in the sound inputs.
However, a software simulation of vowel recognition by Uysal et al [USH06] shows the general robustness of spike-based computation against noise. For each channel of a software cochlea based on Meddis Hair Cell Model, they evaluated the synchrony in the ISI and applied rank order coding for classification. The percentage of correctly identified vowels in noise was compared with the accuracy of a conventional classifier for small training dataset, a 1-nearest neighbour classifier (1NN). Results showed that "both algorithms perform similarly under high SNR conditions" [USH06], however, at lower SNRs as e.g. 5 dB, "the presented implementation outperforms the conventional method by 10.1 % for pink and 19.1 % for white noise.

To sum up, experiments indeed indicate that, after some years of further improvement, event-based silicon cochleae can become an alternative to the established systems of speech recognition and speaker identification. Moreover, the biomorphic sensors seem to have the potential of offering a greater robustness against noise. Though, the studies presented before do not address the issue of whether there is also an advantage concerning computational cost and power consumption.
Some information was however given for an improved and slightly extended version of AEREAR2 by Liu et al [LvSMD14]. For a sound localization task, the computational cost was calculated to be up to 40 times lower than for a conventional system based on cross-correlation as the latter samples the audio signal with high frequency and precision, while event-based "computation is driven by signal activity." [LvSMD14]. Regarding power consumption, the authors do not give numbers, but also expect advantages compared to conventional systems which seems plausible as the computational cost is considerably lower.

Chapter 4

Conclusion and future work

Event-based neuromorphic silicon cochlea chips present themselves as an interesting alternative to conventional speech recognition and speaker identification systems as their performance is in general promising and they offer greater noise robustness and reduction of computational cost, although still more experiments need to be done to test for performance in realistic, e.g. noisy, environments or with accented speech. However, the use of classifiers based on machine learning could be a good approach also to deal with variable speaking styles as these systems can adapt their processing to different, challenging inputs with time. In previous applications, Support Vector Machines (SVMs) were used, which are restricted to inputs with fixed dimensions. Here, improving pre-processing in order to reduce the amount of data that is lost in the reduction process can lead to better results.
Experiments could also be done with classifiers based on artificial neural networks to test for possible advantages.
Orienting more closely towards biology, Abdollahi and Liu [AL11] plan to extend their digit recognition system with a spike-based multi-neuron learning chip, consisting of a VLSI network of integrate-and-fire neurons and plastic synapses, that works in real-time with biologically realistic time constants (refer to [MFI09] for more information). In general, as nature provides an optimal solution for the problem of sound analysis, mimicking it as far as possible is a plausible approach, although this goes with having to understand and build very complex systems.
The neuromorphic concept directly leads to another potential application, namely the replacement of damaged biological cochleae in humans by neuromorphic cochlea chips. The state-of-the-art systems for deaf people are cochlea implants that at least allow for understanding speech, but the sound is perceived as artificial, e.g. the perception of music differs enormously if compared to healthy listeners; additionally, the devices are still large, power-demanding and as usually not fully implantable, there is the risk of damaging the external part.
However, event-based neuromorphic silicon cochlear chips could allow for some improvements. As they are highly integrated, completely implantable devices are conceivable, which leads to stringent requirements concerning power consumption which

could indeed be met: For example, Sit and Sarpeshkar [SS08] succeeded in developing an analog cochlear implant processor that only consumes 357 µW; digital solutions often need 5 mW or more. Additionally, their system can preserve fine-phase-timing information of the signal and therefore significantly improve music perception. Also a promising neuromorphic cochlear implant test chip "including necessary electronics for patient testing has been implemented" [Mar07], providing a digital programming interface to make patient adaptation possible.
But still no fully implantable device is existing. Such a system could use implantable microphones put directly behind the eardrums and, with sufficiently reduced power consumption, batteries that can be charged overnight by inductive coupling. A digital control unit should offer the possibility to adjust the software to the patient and allow for the transmission of further improved algorithms.

At the moment, regardless of the later application, silicon cochleae themselves could still be improved in robustness against the unavoidable variances in analog elements, as this is crucial to e.g. match the responses between channels as close as possible. Additionally, delay must be kept very small to be able to implement real-time-systems, but at the same time a high spectral selectivity is desirable. Finally, in comparison to biology, the input dynamic range of all developed silicon cochleae is yet smaller by several orders of magnitude.

To conclude, although promising, there is still a way to go until neuromorphic auditory sensors with silicon cochleae can be used for speech recognition in smart home applications, as at the moment, e.g. a speaker identification rate of 92 % is not enough to use the system in critical areas as mobile banking or telemedicine. Until now, the system is also not able to replace standard cochlear implants for deaf people. But as it is object of perpetual research and development, considerable improvements and landmark inventions will only be a matter of time.

List of Figures

Bibliography

[AL11] Mohammad Abdollahi and Shih-Chii Liu. Speaker-independent isolated digit recognition using an aer silicon cochlea. In *Biomedical circuits and systems conference (biocas), 2011 ieee*, pages 269–272. IEEE, 2011.

[CLVS07] Vincent Chan, Shih-Chii Liu, and André Van Schaik. AER EAR: A matched silicon cochlea pair with address event representation interface. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 54(1):48–59, 2007.

[Fra98] E. Fragnière. *Analogue VLSI Emulation of the Cochlea*. PhD thesis, Ecole Polytechnique Federale de Lausanne, 1998.

[GY08] Mark Gales and Steve Young. The application of hidden markov models in speech recognition. *Foundations and trends in signal processing*, 1(3):195–304, 2008.

[Ham08] Tara Julia Hamilton. *Analogue VLSI Implementations of Two Dimensional, Nonlinear, Active Cochlea Models*. PhD thesis, The University of Sydney, Australia, 2008.

[Kin05] Peter R Kinget. Device mismatch and tradeoffs in the design of analog circuits. *Solid-State Circuits, IEEE Journal of*, 40(6):1212–1224, 2005.

[LD10] Shih-Chii Liu and Tobi Delbruck. Neuromorphic sensory systems. *Current opinion in neurobiology*, 20(3):288–295, 2010.

[LD+15] S. Liu, T. Delbruck, et al. *Silicon Cochleas*, chapter 4, pages 71–90. John Wiley & Sons, Ltd., 2015.

[LDL12] Cheng-Han Li, Tobi Delbruck, and Shih-Chii Liu. Real-time speaker identification using the aerear2 event-based silicon cochlea. In *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*, pages 1159–1162. IEEE, 2012.

[LM88] Richard F. Lyon and Carver. Mead. An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1119–1134, 1988.

[LVSMD10] Shih-Chii Liu, André Van Schaik, Bradley A Minch, and Tobi Delbruck. Event-based 64-channel binaural silicon cochlea with q enhancement mechanisms. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 2027–2030. IEEE, 2010.

[LvSMD14] Shih-Chii Liu, Andre van Schaik, Bradley A Minch, and Tobi Delbruck. Asynchronous binaural spatial audition sensor with 2 64 4 channel output. *Biomedical Circuits and Systems, IEEE Transactions on*, 8(4):453–464, 2014.

[LW05] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on hmm and svm. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 8, pages 4898–4901. IEEE, 2005.

[Lyo90] RF. Lyon. *The Mechanics and Biophysics of Hearing. Lecture Notes in Biomathematics*, volume 87, chapter Automatic gain control in cochlear mechanics, pages 395–420. Springer, New York, 1990.

[Mar07] Jan-Tore Marienborg. *Neuromorphic Cochlear Implant*. PhD thesis, University of Oslo, 2007.

[MFI09] Srinjoy Mitra, Stefano Fusi, and Giacomo Indiveri. Real-time classification of complex patterns using spike-based learning in neuromorphic vlsi. *Biomedical Circuits and Systems, IEEE Transactions on*, 3(1):32–42, 2009.

[Shi04] H. Shirashi. *Design of an Analog VLSI Cochlea*. PhD thesis, School of Electrical and Information Engineering, University of Sydney, Australia, 2004.

[SLM98] Rahul Sarpeshkar, Richard F Lyon, and Carver Mead. A low-power wide-dynamic-range analog VLSI cochlea. In *Neuromorphic systems engineering*, pages 49–103. Springer, 1998.

[SS08] Ji-Jon Sit and Rahul Sarpeshkar. A cochlear-implant processor for encoding music and lowering stimulation power. *IEEE Pervasive Computing*, 7(1):40–48, 2008.

[USH06] Ismail Uysal, Harsha Sathyendra, and John G Harris. A biologically plausible system approach for noise robust vowel recognition. In *Circuits and Systems, 2006. MWSCAS'06. 49th IEEE International Midwest Symposium on*, volume 1, pages 245–249. IEEE, 2006.

[VOR16] Anup Vanarse, Adam Osseiran, and Alexander Rassau. A review of current neuromorphic approaches for vision, auditory, and olfactory sensors. *Frontiers in neuroscience*, 10, 2016.

[vSF+96] A. van Schaik, E. Fragniere, et al. Improved silicon cochlea using compatible lateral bipolar transistors. *Advances in Neural Informations Processing Systems*, pages 671–677, 1996.

[WB09] Bo Wen and Kwabena Boahen. A silicon cochlea with active coupling. *IEEE Transactions on Biomedical Circuits and Systems*, 3(6):444–455, 2009.

[WB11] Bo Wen and Kwabena Boahen. *Integrated Microsystems: Electronics, Photonics, and Biotechnology*, chapter 10 A Biomorphic Active Cochlear Model In Silico, pages 207–235. CRC Press, 2011.

[WK+92] L. Watts, D. A. Kerns, et al. Improved implementation of the silicon cochlea. *IEEE Journal*, 27(5):692–700, 1992.

License