Speech Coding and Information Processing in the Peripheral Human Auditory System

Huan Wang

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Realzeit-Computersysteme

Speech Coding and Information Processing in the Peripheral Human Auditory System

Huan Wang

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender:	Univ. Prof. Dr.-Ing. habil. Gerhard Rigoll
Prüfer der Dissertation:	1. Univ. Prof. Dr.-Ing. Werner Hemmert
	2. Univ. Prof. Dr.-Ing. Georg Färber

Die Dissertation wurde am **05.01.2010** bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am **26.04.2010** angenommen.

Acknowledgements

I would like to express my greatest appreciation to Prof. Dr. Werner Hemmert, who has been guiding me during the whole period of this PHD dissertation. Without his trust, his encouragement and support, I could never have reached so far. I am deeply indebt to his commitment to me and to my research effort. He has taught me so much in doing scientific research. And he spent days and nights with me editing the manuscripts that I have written, word by word, sentence by sentence. Werner is not only an excellent "Doktorvater" but also a very valueable friend. I am greatly thankful for the friendship with him, and his family.

I sincerely thank Prof. Dr. Georg Färber for his kindness to supervise and review my work. I highly appreciate those helpful discussions with him and the seminars organized by him in the institute. He has set for me an excellent example of an outstanding researcher and a respectable person.

I would like to thank my colleagues in our research lab, Marcus Holmberg, Mathias Mamsch, Cuchi Hernandez Francisco, Michael Isik for the teamwork that we have done, for the friendship that we have built, and for the nice times that we have spent together. My gratitude also goes to David Gelbart for his help and cooperation. His insight in speech recognition technology has been an indispensible contribution to our research work. I am very thankful to many colleagues and friends in Infineon Technologies AG, for their support, trust, discussion, interesting activities together and the big farewell party.

I am beholden to my family and my friends in China and in Munich. Their company, care, support and love give me the confidence to persue my dreams, be it the dissertation or whatever else in life. My parents supported me and encouraged me any time that I needed, without them this PHD work wouldn't have been possible at all.

My special thanks go to my wife Meng Zhang for her never-ending love, friendship and support. Life has become a wonderful journey with her by my side, and I know this journey will be even more exciting in the future.

This work was carried out while I was employed at Infineon Technologies AG, Munich, Germany. It was funded by the German Federal Ministry of Education and Research (BCCN Munich, reference numbers 01GQ0441 and 01GQ0443).

Huan Wang

München, September, 30th, 2009

To Meng and Emma

Contents

List of Figures

List of Tables

List of Symbols

RCS	Lehrstuhl für Realzeit–Computersysteme
AM	Amplitude Modulation
AN	Auditory Nerve
ANF	Auditory Nerve Fiber
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BM	Basilar Membrane
CF	Characteristic Frequency
DCT	Discrete Cosine Transform
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
HRTF	Head Related Transfer Function
HTK	Hidden Markov Model Toolkit
IHC	Inner Hair Cell
MFCC	Mel Frequency Cepstrum Coefficients
MI	Mutual Information
MLP	Multi-layer Perceptrons
MTF	Modulation Transfer Function
OA	Offset Adaptation
ON	Onset Neuron
PSTH	Post Stimulus Time Histogramm
roex-function	rounded exponential function
SII	Speech Intelligibility Index
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
WER	Word Error Rate

Abstract

The human auditory system outperforms existing automatic speech recognition system by a large extent. This motivates the modeling of the human auditory system and the application of auditory model for robust automatic speech recognition. During the last years, we have aquired a good knowledge about the mechanisms of the auditory system. However, we are less sure about which properties of the auditory coding contributes to its outstanding performance. The debate about different coding strategies such as "rate code" or "temporal code" still persists. This thesis presents a framework including auditory modeling, speech recognition and information theory, in search of the essential aspects of auditory coding.

I used a detailed model of human auditory processing that generates realistic auditory nerve spike trains. The model includes strong compression and level dependent shape of active inner ear filtering. I augmented the model with a realistic offset adaptation mechanism to replicate the "dead-time" period of the auditory nerve fibers after signal onset. I also modeled cochlea nucleus onset neurons and connected them to the model.

I tested the model qualitatively using a realistic scenario: speech in noise. I implemented different speech recognition engines based on the HTK testbed and the MLP testbed. I evaluated the output auditory nerve spike trains using a rate-place based coding strategy and then augmented it with spikes from octopus neurons which codes mainly the temporal structure of the speech signal. Results showed that the realistic offset adaptation improves temporal coding of auditory nerve fibers which also leads to considerable improvement in recognition results. The model achieved robustness against noise which is similar to human performance, even though there is still a big gap in absolute performance. The MLP testbed outperformed HTK testbed, which suggests that innovative features such as neural spike trains require proper matching to the back-end.

I also implemented an approach based on information theory for quantitative assessment of the auditory model. I found out that information saturates for periodic signals such as spoken vowels. I was able to analyze the impact of temporal resolution of the spike trains from different onset neurons. I suggested that onset neurons utilized a very high temporal resolution to code information. The quantitative measurement also enabled me to analyze the effect of noise on information transmission.

In conclusion, the human auditory model, the qualitative measurement using speech recognition and the quantitative measurement using information theory complement each other. Together, they provide a framework for continuously improving the performance of human auditory modeling for the purpose of speech recognition and information processing.

1 Introduction

Nowadays, people are surrounded by more and more electronic devices, which require interfaces to enable human machine interaction. Life would have been much easier, if we were able to talk to the devices as the astronauts do to the supercomputer "Hal" in the movie *2001: A Space Odyssey*.

Voice commanded applications, according to market analysis, are expected to cover many aspects in our future life. The most likely candidates range from general applications with computers, telephones, call centers to many other specific areas such as car navigation system, hand held devices, for which keyboards are less plausible, as well as many services where acquisition of initial information can be done by computers in order to reduce the cost. Speech recognition is currently regarded by the market as one of the most promising technologies of the future. A 50 fold rise in sale of industrial speech technology products was seen from the year of 1997 to the year of 2003 Becchetti and Ricotti [2002].

If speech recognition by machine were a solved problem, then work in this field would only consist of the application of proven technology to new tasks in the aforementioned areas, and basic research would not be necessary. This is to some extent almost the current situation. Three decades of intensive research in speech recognition has led to efficient algorithms in feature extraction; Statistical approaches such as Hidden Markov Models (HMM) and Multilayer Perceptrons (MLP) have been developed to the level where excellent performance can be achieved in the laboratories on large vocabulary continuous speech recognition tasks. Commercial products with innovative applications have been around for many years, slowly gathering momentum in the market.

However, speech recognition is by any fundamental sense *not* a solved problem. Current speech recognition works well in laboratory environment but breaks down drastically as soon as noise or reverberation distorts the speech signal, not to mention that in reality speech recognition systems have to deal with various accents, imperfect grammar, unfamiliar words, etc. On the other hand, when it comes to speech interaction with machines, users become less tolerant to speech recognition system than they do to the keyboards. Users expect speech recognition system to behave as humans do. However, current commercial products are still striving to achieve a balance between recognition performance and the complexity of the system. One of the solutions was to limit the size of the vocabulary and the complexity of the dialog system to achieve acceptable recognition performance for a specific application.

Problems stated, it is however very difficult to find the solutions based on existing techniques. Both the feature extraction (also referred to as "the front end") and the recog-

nition engine (also referred to as "the back end") are genuinely problematic: Feature extraction schemes based on frequency transformation drop most of the essential temporal information, which are believed to be instrumental for robust speech recognition [Delgutte and Kiang, 1984a, Silkes and Geisler, 1991, Sachs and Young, 1979]; The statistical approach (HMM) used to perform recognition tasks make numerous assumptions about speech, some of which obviously unrealistic [Morgan and Bourlard, 1995].

In comparison to automatic speech recognition systems, humans possess stunning capability in speech perception. It is the most robust speech recognition system that we are aware of so far. The performance of human auditory system dwarfs today's most sophisticated computer system. It is therefore least surprising that researchers have been trying to improve speech recognition systems based on principles motivated by the human auditory system. A natural approach to reach robustness, or at least to understand the mechanisms behind it, is to mimic the peripheral human auditory system.

The state-of-the-art speech processing technology already utilizes many principles motivated by human auditory processing. von Helmholtz [von Helmholtz, 1863] put forward the place-coding as the principle mechanism of hearing. His notion that human ear analyzes individual frequency components of sound signals has motivated the widespread use of magnitude spectra in sound processing technologies. Specific algorithms such as Mel frequency filter banks, log energy could all easily find their counterparts in human auditory system. Multilayer Perceptrons were also developed using some "organizational" principles believed to be used in the human brain, with the hope that the network will possess some of the brain's desirable characteristics such as massive parallelism, learning ability, generalization ability etc. However, human auditory system differs from current automatic sound processing system in a fundamental way, i.e, human auditory system codes sound signal into trains of action potentials of the neurons, which are transmitted to higher processing stages of the brain.

In order to better mimic the human auditory processing, we have developed a phenomenologically motivated inner ear model. Our model is based on the latest biological and neurological findings, and very nicely replicates the physiological data. Such a realistic model is meaningful in many ways. First, it casts light on the mechanisms used by auditory processing. This is straightforward given that our model is built to resemble the human auditory processing. Therefore, every improvement in the modeling helps us better understand sound processing by humans or other mammals. Second, a suitable model makes it possible for researchers to generate data which is at least realistic to some extent, and to study the property of auditory processing. Of course such a model can not substitute physiological or neurological experiments. But it serves as a complementary approach of auditory study and makes researchers less dependent on measured data, which is limited in amount and also expensive to get access to. Last but not the least, it is possible to develop practical applications with our inner ear model, such as sound localization or speech recognition systems.

Besides the efforts of modeling human inner ear and constantly improving the performance of the model based on physiological findings, this thesis also covers the topic of applying

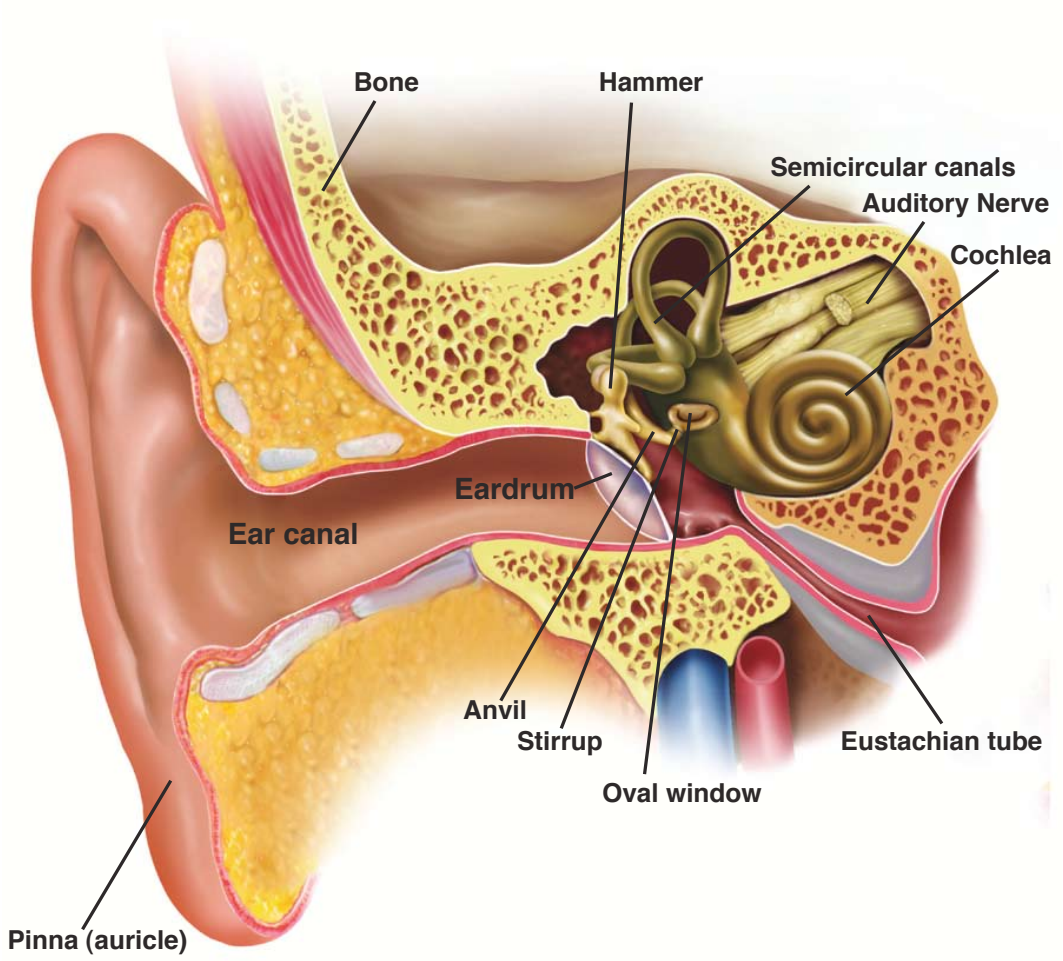

Figure 1.1: Overview of the hearing organ.

the model to perform speech recognition tasks. We successfully integrated our inner ear model as the front end into existing speech recognition frameworks and achieved comparable results to standard feature extraction approaches applied in commercial speech recognition systems. Another major part of the work comprises of analyzing the auditory model based on information theory. Compared to automatic speech recognition, information theory based approach makes not only less modification of the output from our inner ear model but also less assumption about the underlying processing. Therefore it gives a more intuitive and rigorous assessment of the auditory processing.

1.1 Human Auditory System

Human auditory system is one of the most complicated sensory system known to us. Figure 1.1 gives an overview of the auditory system. The pinna (auricle) gathers the sound energy, which is lead through the outer ear canal to the ear drum (tympanic

membrane). The pinna plays an important role in our ability to distinguish the direction of a sound source, especially in the elevation angle [Blauert, 1974, 1997]. The outer ear canal acts like an open pipe with a length of 20 to 30 mm. Its resonance frequency lies around 4 kHz. It is responsible for the exceptionally high sensitivity of hearing system in this frequency range.

The middle ear transforms the sound signal from air particle movement in front of the ear drum to fluid motion in the inner ear. The transformation is performed over a wide frequency range. The middle ear is responsible for matching the impedance in order to minimize energy loss due to reflection, which means that the maximal amount of energy is transferred to the inner ear. The sound pressure in the outer ear canal sets the eardrum into vibrations, which are then transmitted via the hammer (malleus) and the anvil (incus) to the stirrup (stapes). The stapes' foot plate contacts the oval window, a hole forming the entrance to the inner ear. The Eustachian tube connects the air filled middle ear cavities to the upper throat. It opens and closes periodically, keeping the static pressure in the cavity at atmospheric pressure level. In summary, the pressure of the sound signal is transduced into displacement of the oval window at the entrance of the inner ear.

In the inner ear, the transformation from mechanical entities into nerve action potentials takes place. This is one of the most important stage of the hearing process. Cues of the sound signals have to be coded for neural processing, and any information loss will not be available for higher processing stages.

The inner ear is constituted of the cochlea, the semicircular canals and the organ of balance, among which only the cochlea contributes to hearing. The cochlea is a spiral formed, fluid-filled organ in the temporal bone. It is coupled to the middle ear through the round window and the oval window, which locate on the opposite site of the cochlea partition. The footplate of the stapes is inserted into the oval window, whereas the round window is covered by a membrane. The modiolus forms the central axis of the cochlea through which the auditory nerve fibers run. Figure 1.2 shows a schematic cross sectional view of the human cochlea. The three ducts – scala vestibuli, scala media, and scala tympani – spiral two and a half turns around the modiolus. The cochlear partition, which separates the scala media and scala tympani, is of great interest to us. The cochlear partition consists of a thin shelf of bone projected from the modiolus (the osseous spiral lamina) and the basilar membrane, which is attached to the osseous spiral lamina and to the outer wall of the cochlea through the spiral ligament. At the apical tip of the cochlea, there is a narrow opening in the membrane, known as the helicotrema, which constitutes a connection between scala vestibuli and scala tympani. The basilar membrane (BM) mainly consists of extracellular material. Its stiffness decreases gradually towards the apex. In the base of the cochlea the membrane is narrow and thick, whereas at the apex it is wide and thin.

When the stapes moves into the cochlea, it causes fluid motion. The fluid pushed on the cochlear partition, and because of the pressure equalizing properties of the round window, the partition is set in motion. The motion travels from the basal end of the cochlea to the apex, generating the so-called traveling wave, whose propagation is determined by the

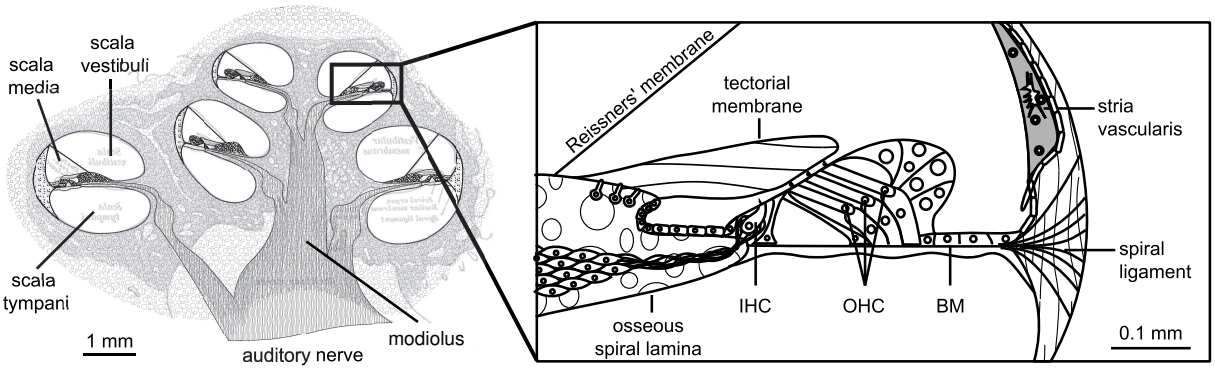

Figure 1.2: Cross section of the cochlea. Basilar Membrane (BM), inner hair cell (IHC) and outer hair cell (OHC).

mechanical properties of the cochlear partition and the surrounding fluid moving with it.

The stiffness of the cochlear partition decreases from the oval window to the apex of the basilar membrane. The cochlear partition also shows damping effects and inertia, mainly from the surrounding fluids moving with the membrane. All together the cochlear partition exhibits a resonance behavior, with a resonance frequency that grades from base to apex. A resonance frequency corresponds to each location along the cochlea (the characteristic location of the frequency). A wave with a frequency that is the same as the resonance frequency (characteristic frequency, or CF) reaches its maximum amplitude, looses its energy and quickly diminishes. Waves with lower frequency will propagate further. This is the famous "traveling wave" model, as measured by Békésy [1960]. The traveling wave model is the foundation of the tonotopic decomposition of the incoming sound signal.

The mechanical signal is then processed by auditory nerve fibers into discrete spike trains, which are transmitted further to the brain for neuronal processing. A detailed explanation to the mechanism of human hearing system can be found in Section 1.4.

1.2 Speech Recognition

The highly complex auditory system serves several critical missions, among which speech perception is probably the most important one. Therefore it is only logical that we chose speech recognition as a criterion to evaluate our auditory model qualitatively.

Automatic speech recognition system shares the basic procedures and mechanism of human beings. In essence, ASR systems, like human beings, rely on a parametric model where knowledge can be gained by parameter tuning. Parameters are set to recognize the model which represents a certain phenomena in the most accurate way [Becchetti and Ricotti, 2002]. Human beings process information with very large biological neural networks where the parameters are stored in the connections, while ASR implements much

simpler Hidden-Markov-Models or artificial neural networks.

There are several basic steps involving in an ASR system: initialization, training and recognition. At the beginning, the ASR system must be initialized. For ASR using based on HMM, initialization includes defining the HMMs' structures and estimating the starting values of HMM parameters. In a second step referred to as "training", the model gains knowledge by repeatedly "observing" examples of the objects to recognize, i.e. acoustic sounds and their associated meaning (transcripts). The larger the number of repetitions of training, the more accurately the model reflects real examples of speech. The training process resembles the early ages of human beings.

In the recognition step of the ASR, strings of words/alphabets are associated to speech. Recognition is performed seeking the string of words/alphabets that best match the acoustic observations based on the available model. For humans, recognition consists of associating speech to the related concepts and it is reasonable to assume that the brain performs a similar procedure as the ASR system by searching for the best match.

The recognition accuracy basically depends on the quantity of the training material observed by the model and by the capability of the model to continuously "absorb" knowledge, i.e. progressive training or adaptation, etc. Human brains fundamentally outperform ASR systems in two aspects: the models of ASR systems are far simpler than the part of the brain dedicated to speech recognition; human brains learn new knowledge years by years and adapt to new scenarios continuously, an ability which the ASR system can't even get close to. For this reason, ASR system only provides a primitive analog to the human brain. Some experts argue that ASR performance maybe comparable to that of a two-year-old child [Comerford et al., 1997]. Nevertheless, ASR system provides an efficient tool for qualitatively assessing the coding strategy of the modeled auditory system, compared to methods that only look into certain well-defined properties of speech, such as formants [Conley and Keilson, 1995]. Furthermore, the training procedures of ASR systems are improving, training material is increasing, and more complex HMM models are becoming feasible. Thus in future years, ASR will definitely grow "older and older" in performance with respect to the capabilities of a two-year-old child.

1.3 Information Processing

We used information theory[Cover and Thomas, 1991a] to quantitatively analyze the auditory system, as a complement to the qualitative approach of automatic speech recognition. Automatic speech recognition is a meaningful but indirect indicator in evaluating the auditory system, mainly because not only the auditory system, but also the recognizer could influence the performance. At the mean time, the largely complex recognition system prevents any direct quantitative conclusion. Another limitation of the speech recognition system is that it operates at low temporal resolution (usually 10 ms). Therefore it fails to provide insight to the temporal property of the auditory system, which utilizes high dimensional and sparse spike trains which are accurate in time to transfer information.

Information theory plays an important role in this part of my thesis. At the methodological level, we used the important quantities of information theory – such as entropy, mutual information, and redundancy – to quantify the properties of the stochastic neural activity. More importantly, information theory provides a conceptual framework for thinking about principles of neural activities. The search for design principles that govern the processing of sensory cells, was boosted by the appearance of Shannon's information theory. Attneave [1954] suggested analogies between sensory systems and communication channels, which gave the ground for postulating optimization principles for neural circuits. Several researchers discussed generic principles that could underlie sensory processing [Barlow, 1959a, Atick, 1992, Becker, 1996], among which the information maximization and redundancy are worth to be mentioned here.

The information maximization principle (InfoMax) was put forward by Linsker [Linsker, 1988, 1989], suggesting that a neural network should tune its circuits to maximize the mutual information between its outputs and inputs. Since the network usually has a prefixed architecture, this leads to a constrained optimization problem for any given set of inputs.

Redundancies in sensory stimuli were put forward as important for understanding perception since the very early days of information theory. Barlow's specific hypothesis [Attneave, 1954, Barlow, 1959a,b] was that one of the goals of a neural system is to obtain an efficient representation of the sensory inputs, by compressing its inputs to achieve a parsimonious code. During this compression process, statistical redundancies that are abundant in natural data are filtered out such that the neural outputs become statistically independent. This principle was hence named Redundancy Reduction.

Barlow further suggested [Barlow, 2001] that the actual goal of the system is rather redundancy exploitation, a process during which the statistical structures in the inputs are removed in a way that reflects the fact of how the system uses it to identify meaningful objects and structure in the input. These structures are later represented in higher processing levels, a process that again yields a reduction in coding redundancies of higher level elements. For example in speech signals, pitch frequency is one of the meaningful structures as mentioned above.

1.3.1 Representation of Neural Signals

Estimating mutual information (MI) from empirical distributions is a difficult task, in particular with the small sizes of data available. A naive approach to this problem would be to estimate the joint distribution of stimuli versus all possible neural responses, and then to estimate the mutual information of this high dimensional distribution. Unfortunately this approach is always bound to fail due to the richness of neural responses. For example, a typical pyramidal neuron in the cortex fires spikes that should be measured with a temporal resolution of 1-4 ms [Singer and Gray, 1995], and can thus produce in theory at least 2^{250} different spike trains in a single second. Since a robust estimation of

a probability density function requires obtaining many samples relative to the number of possible responses, this approach is doomed to fail.

The crucial observation is that MI estimation does not require estimating the full joint distribution of stimuli and responses. One reason is that the set of functionally distinct neural responses is much smaller. And the other reason is that mutual information is a scalar function of the distribution, which actually averages the log-likelihood ratio $\log \frac{p(x,y)}{p(x)p(y)}$ over all $x's$ and $y's$. Its estimation is therefore expected to be more robust than the estimation of the distribution itself [Nemenman et al., 2002].

The estimation of MI from a finite sample involves an important tradeoff between the reliability of estimation and model complexity with which we represent the output of the neurons. If we could represent the neural output with a low dimensional signal, it is obviously easier to get an accurate estimation of its probabilistic density function and thus a better estimation of the MI. However, any representation of the neural output which reduces the dimensionality leads to a loss of mutual information as indicated by the data processing inequality [Cover and Thomas, 1991b].

The challenge in MI estimation is therefore to find low complexity representations of spike trains that are still highly informative. This makes it possible to obtain both a high level description and a reliable estimation of the information they convey. Therefore we will try to reduce the dimensionality of the experimental data for the calculation of mutual information.

In practice, a lot of techniques have been developed to represent neural responses with low dimensionality, each focusing on a different aspect of spike trains. These specific methods for transforming spike trains into low dimensional representations are [Chechik, 2003]:

1. Spike counts
2. Spike counts weighted by inter-spike-intervals
3. First spike latency
4. Spike patterns as binary words (direct method)
5. Legendre polynomial embedding
6. Second order correlation between spikes

Among these six different methods, the direct method and first spike latency achieve the maximum information [Chechik, 2003]. In this thesis, we used the direct method to calculate mutual information. We temporally downsampled the spike trains (digital form) and used time to represent each spike, a process that we later referred to as "binning". We used this low dimensional representation of neural responses to keep the accurate information about spike timing, as the timing of individual spikes may represent the basic mechanism of information transmission along the neural pathway. Theoretically, this representation will perfectly preserve the information carried by spike trains if the time we used to represent spikes is continuous (in this case, the dimensionality does not

decrease at all). In our calculation, we used various temporal resolutions to represent spike trains. Therefore, the dimensionality of spike trains is reduced, and as a result, part of the information gets lost.

This representation approach is essential in evaluating temporal properties of the neural system. Input and output of the neural system can be represented with variable temporal resolutions. By comparing the mutual information that was transmitted using each temporal resolutions, we gain insight about the temporal properties of the neural system.

1.4 Model of the peripheral auditory system

Previoisly, our group has developed an inner ear model which codes audio signals into auditory nerve action potentials. The details of the modeling is described in detail in Holmberg [2007], Hemmert and Holmberg [2009]. This section provides a brief introduction to the model. Improvements which were introduced for this thesis are described in Chapter 2.

1.4.1 Outer ear, ear canal and middle ear model

The model includes a simple joint model of outer ear and ear canal. The model is implemented as a transfer function from free field sound pressure to sound pressure at the ear drum. It is closely related to the head related transfer function (HRTF), which is very important for sound localization [Blauert, 1997]. The transfer function simulates the pinna effects and ear canal resonance which mainly amplifies frequencies around the second and third formants. Therefore, including the outer ear and ear canal has important implications for how well the speech spectrum is encoded in the auditory nerves. In Chapter 4 and Chapter 5 where we performed ASR tests and information calculation, we included the outer ear and ear canal model. However in this section where the model output was compared with physiologic measurements, the outer ear and ear canal was not included, since physiological measurements usually report responses to stimuli presented directly at the ear drum.

A high-pass filter (first-order filter, corner frequency: $f_c = 1$ kHz) mimics the transformation of sound pressure into the vibration velocity of the ossicular chain. Modeling the middle ear transfer function as a high-pass function, rather than a band-pass function which is usually the case, was motivated by the findings of Ruggero and Temchin [2002, 2003]. Ruggero and Temchin [2003] argued that the high-frequency cut-off of the living middle ear transfer function occurs at frequencies above what is relevant for the audiogram. Thus, it has only limited effect on the cochlear frequency analysis and was not included in the present model.

1.4.2 Inner ear hydrodynamics

The hydrodynamical inner ear model is based on a transmission-line model of the inner ear [Strube, 1985]. The transmission-line provides a one-dimensional solution of the passive inner ear hydrodynamics [Peterson and Bogert, 1950, Oettinger and Hauser, 1961, de Boer, 1980, 1984]. The electrical equivalent circuit of the mechanical model (using a mobility type acoustical analogy) is shown in Figure 1.3b. Each resonator i models the damped mass-spring-system of a portion of the basilar membrane. The model is discretized in equidistant spaced intervals Δx along the length of the cochlea. The distance from the stapes is denoted by x. Any parameter Z described below takes a value $Z_i = Z(i \cdot \Delta x)$ as a result of the descretization. A detailed physical interpretation of the electrical entities in Figure 1.3 can be found in Strube [1985]. Here it sufficed to simply state that the inductances L_i are linked to acoustic mass, the capacitors C_i to volume compliance, and the resistances R_i to acoustic resistances of a section of the fluid-loaded basilar membrane. L_{ci} describes the coupling of adjacent cochlea sections by fluid mass (for a long-wave model) and consequently varies with the area of the cochlear duct. Following the notation of Strube [1985], the other variables in Figure 1.3 have the following interpretations:

J_0 volume velocity driving the inner ear

I_i transversal basilar membrane volume velocity

J_i longitudinal volume velocity of inner ear liquid

The purpose was not to do a parameter study, but to achieve pre-defined resonant frequencies and filter bandwidths. The acoustic mass, and thus the inductance $L(x)$, was modeled to vary according to an exponential map (see Strube, 1985 and Viergever, 1980 for details)

$$L(x) = \frac{L_0}{\Delta x} \cdot e^{(L_x \cdot x)} \tag{1.1}$$

The other free parameters are determined by maps of the desired resonant frequency and quality factors. If the second-order resonators were unconnected, the compliance $C(x)$ could be calculated by

$$C(x) = \frac{1}{(2\pi f_{res}(x))^2} \cdot \frac{1}{L(x)}, \tag{1.2}$$

where $f_{res}(x)$ describes a mapping between resonant frequency and place. The interconnection of sections however changes this relation. The value of $C(x)$ was therefore determined by an iterative procedure described below. The resistances can be calculated from the quality factor Q (again for unconnected resonators)

$$R(x) = \sqrt{\frac{L(x)}{C(x)}}/Q(x) \tag{1.3}$$

To adjust the hydrodynamics model to given maps $f_{res\,hyd}(x)$ and $Q_{hyd}(x)$ (compare Sec. 1.4.4), Eqs. 1.2 and 1.3 were used to calculate starting values. In five iteration steps,

the resonant frequency of each location x was determined from the Fourier-transformed impulse response. The value of $C(x)$ was updated, and $R(x)$ calculated from Equation 1.3. The resulting frequency map differed less than 3% from the given specification at all frequencies above 150 Hz.

In a lossless cylindrical tube, the impedance of each section is described by a purely inductive element, or mass term. Assuming a slow-varying cross-section area $A_{scalae}(x)$, then

$$L_c(x) = \frac{2\rho\Delta x}{A_{scalae}(x)} \tag{1.4}$$

where ρ is the density of the scala fluids. The duct area is a logarithmic fit to a measurement of the human cochlea scala areas [Thorne et al., 1999][1)].

$$A_{scalae}(x) = A_{scalae\,0} \cdot e^{A_{scalae\,x} \cdot x} \tag{1.5}$$

The helicotrema, a hole at the most apical part of the BM (Figure 1.3a), prevents the buildup of static pressure across the BM and was modeled as an acoustic tube (L_h, R_h, Dallos, 1970)

$$Z_h = R_h + j\omega L_h = \frac{8\eta l_h}{\pi \cdot r_h^4} + j\omega \frac{4\rho l_h}{3\pi \cdot r_h^2} \tag{1.6}$$

where r_h is the radius of the helicotrema and l_h the length of the acoustic tube. The volume velocity in front of the stapes, modeled as an ideal current source, drives the model:

$$J_0(t) = v_{stapes}(t) \cdot A_{stapes} \tag{1.7}$$

The displacement of the basilar membrane, $x_{BM\,hyd}(x)$, can be derive form the pressure ($U_C(x,t)$) across each sections' volume compliance $C(x)$

$$x_{BM\,hyd}(x,t) = g_{BM}(x) \frac{U_C(x,t) \cdot C(x)}{\Delta x \cdot w_{BM}(x)} \tag{1.8}$$

where

$$w_{BM}(x) = w_{BM\,0} \cdot e^{w_{BM\,x} \cdot x} \tag{1.9}$$

is the width of the BM and g_{BM} a correction term. The exponential fit to the basilar membrane width was inherited from Viergever [1980] and Strube [1985].

Compression Model

Inner ear hydrodynamics only explains basilar membrane vibrations in a passive (dead) cochlea or in a healthy cochlea at high sound levels. At low levels vibration amplitudes are amplified by as much as 60–80 dB and filter shapes are much sharper [see Robles and Ruggero, 2001 for a recent review]. This section introduces the phenomenological model of the mechanism often referred to as the "cochlear amplifier", which also provides large

[1)] Data is available online at http://oto.wustl.edu/cochlea/mrhmvol.htm

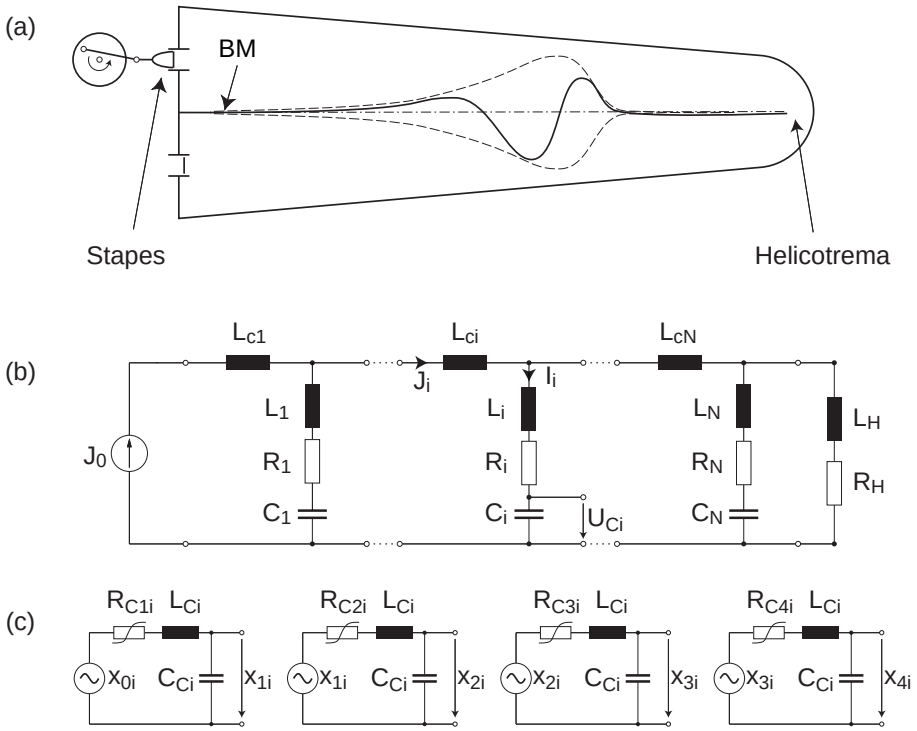

Figure 1.3: (a) Schematic figure of inner ear hydrodynamics. (b) Electrical equivalent circuit (mobility type acoustical analogy). All sections take the form of second-order resonators (L_i, C_i, R_i), which are coupled by fluid inertia L_{ci}. Sound signals are injected by a volume velocity source (J_0) into the transmission line, which is closed by the impedance of the helicotrema (modeled as an acoustic tube, L_h, R_h). (c) The electrical circuit representation of the phenomenological nonlinear "amplification" and compression stage consisting of four resonators showed for one section only. The first stage is driven by BM displacement $x_{0\,i}$ derived from $U_{C\,i}$ in the transmission line model. Three additional stages are driven by the output of the previous stage. The compressed output is available at the last stage $(x_{4\,i})$. All parameters Z take values $Z_i = Z(i \cdot \Delta x)$ as a result of the longitudinal discretization.

dynamic compression. Although the model is purely phenomenological and implemented with time-varying filters, it uses a feedback-loop that is inspired by the functionality of the outer-hair cells.

The basic idea behind the implementation is that the quality factor (Q-factor) of a second-order resonator (Figure 1.3c) provides "amplification"[2)] at the resonant frequency. Varying the Q-factor changes both output amplitude (predominantly close to the resonant frequency, where amplification is proportional to the Q-value) and filter bandwidth; the higher the Q-value (and thus the amplification), the narrower the filter bandwidth. The benefit of using passive resonators instead of an active mechanism, as is probably the case in the living ear, is stability. Many auditory models with compressive nonlinearities use the same general idea, either in models of inner ear hydrodynamics [e.g., Strube, 1985, Deng and Geisler, 1987a] or in banks of auditory-like filters [e.g., Carney, 1993, Meddis et al., 2001, Robert and Eriksson, 1999, Zhang et al., 2001]. The original model of Strube [1985] only changed the Q-factor of the serial resonators R_i, L_i and C_i in Figure 1.3 by varying R_i. To achieve amplification of 60 dB, a Q-factor of 1000 is required, implying an extremely narrow filter with excessive ringing. The model proposed by Strube could therefore only achieve much smaller amplification/compression. To achieve both large amplification and filter shapes as observed in physiological measurements, four[3)] resonators were connected in series to the hydrodynamics model.

The "compression resonators" are shown in Figure 1.3c. The input is displacement calculated for the linear hydrodynamics model ($x_0(x) = x_{BM\,hyd}(x)$). For each location x, the values of $L_C(x)$ and $C_C(x)$ are fixed and equal for all four resonators. Thus the resonant frequency is approximately constant and determined by

$$f_{CFcompr}(x) = \frac{1}{2\pi} \frac{1}{\sqrt{L_C(x) \cdot C_C(x)}}, \tag{1.10}$$

The modulation of the resonator's Q-value is realized by the time varying resistance $R_{Cn}(x,t)$, where n is the resonator number (from 1 to 4). The four resistances $R_{C1}(x,t)$ to $R_{C4}(x,t)$ vary independently from each other and give the resonators Q-values

$$Q_n(x,t) = \sqrt{\frac{L_C(x)}{C_C(x)}} / R_{Cn}(x,t) \tag{1.11}$$

The long-wave approximation to the one-dimensional solution of the inner ear hydrodynamics, which forms the base of the model, is known to have too shallow high-frequency slopes [de Boer, 1996]. The resonators in Figure 1.3c have frequency responses that are flat at low frequencies and have -12 dB/oct slopes well above their resonant frequency.

[2)] Note that "amplification" only means an increased voltage at the resonator's output compared to its input voltage and not power amplification. When referring to "amplification" in this section, we mean this passive amplification of amplitudes.

[3)] Using four resonators is somewhat arbitrary, but provided a good compromise between maximum compression and filter shapes.

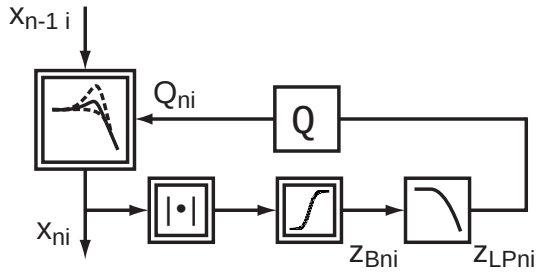

Figure 1.4: Schematics of one "compression resonator" (compare also Figure 1.3c).

Thus, by using such resonators the combined BM model has a below-resonance behavior determined by the hydrodynamics model, an above-resonance slope sharpened by the four resonators, and a level-dependent bandwidth at the resonant frequency.

The schematics of one compression resonator is shown in Figure 1.4. The output of each resonator $x_n(x,t)$ (compare Figure 1.3c) is rectified and passed through a first-order Boltzmann function (normalized to the range $[0,1]$), mimicking the transduction process of the outer hair cells:

$$z_{B\,n}(x,t) = 2 \cdot \left(1 - \frac{1}{1 + e^{-|x_{Cn}(x,t)|/s_n(x)}}\right) \tag{1.12}$$

$s_n(x)$ is a parameter of the model that determines where saturation sets in. The compression model is constructed so that each resonator ($n = 1...4$) compresses a part of the model's total dynamic range. The absolute value of the displacement $|x_{Cn}(x,t)|$ is used in the Boltzmann function, because the quality factor of the resonator should decrease for both positive and negative deflections. $z_{B\,n}(x,t)$ is low-pass filtered to obtain $z_{LP\,n}(x,t)$

$$z_{LP\,n}(x,t) + \tau_{z_B} \frac{\partial\, z_{LP\,n}(x,t)}{\partial t} = z_{B\,n}(x,t) \tag{1.13}$$

(first-order low-pass filter with corner frequency $f_c = 800\,\mathrm{Hz}$). The outer-hair cell membrane properties motivates this step, and in the model it reduces harmonic distortions. Finally, the value is transformed into the instantaneous Q-value of the resonator. The Q-value of each resonator is calculated from

$$Q_n(x,t) = (Q_{max}(x) - Q_{min}(x)) \cdot z_{LP\,n}(x,t) + Q_{min}(x) \tag{1.14}$$

in each time step. The value of $Q_{max}(x)$ determines the maximum amplification and the filter bandwidth at low levels. The ratio of $Q_{max}(x)$ and $Q_{min}(x)$ determines the maximum compression (in dB) as

$$G_{compr}(x) = 20 \log \left(\frac{Q_{max}(x)}{Q_{min}(x)}\right)^4 \tag{1.15}$$

since we used four resonators. The value of $R_{Cn}(x,t)$ finally, is derived from Equation 1.11.

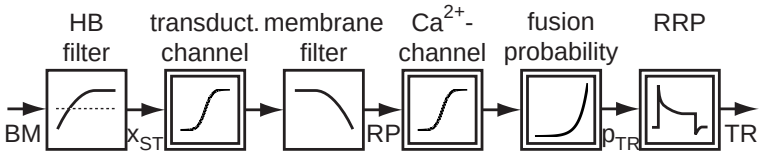

Figure 1.5: Schematics of the inner hair cell model.

1.4.3 Inner Hair Cell Model and Synaptic Mechanisms

The inner-hair cell and the auditory nerve model is adapted from Sumner et al. [2002]. In this section we gave a short description of the model. We also described the modeling of inner hair cell and auditory nerve in Section 2.2 in more detail. A thorough description of the model together with parameters can be found at Sumner et al. [2002].

Figure 1.5 shows the schematics of the model. Basilar membrane vibrations cause fluid motions which in turn drive the hair bundles (HB) of the inner-hair cells (IHCs). As a first approximation, fluid friction and HB stiffness form a first-order high-pass filter, i.e. HB displacement is proportional to BM velocity at low frequencies and to BM displacement at high frequencies. HB motion is scaled due to the geometry of the organ of Corti [Dallos, 2003], which is modeled by introducing a lever gain $g_{HB}(x)$.

$$x_{HB}(x,t) + \tau_{HB}(x)\frac{\partial\, x_{HB}(x,t)}{\partial t} = g_{HB}(x)\cdot\tau_{HB}(x)\frac{\partial\, x_4(x,t)}{\partial t} \tag{1.16}$$

Both lever gain and corner frequency of the filter vary with cochlear location.

The deflection of the hair bundle opens ion-channels at their tips. Channel open probabilities were modeled as a second-order Boltzmann function. Cations (mostly K^+) enter the cell which in turn rises the cell's receptor potential (RP) – the cell membrane depolarizes. The electrical properties of the membrane were modeled by a first order low-pass filter with a corner frequency of approximately 1 kHz [Sumner et al., 2002].

Voltage-dependent calcium channels are located close to the cell's synaptic terminals. Ca^{2+}-channel activation follows a first order Boltzmann function. Elevated Ca^{2+}-concentration causes transmitter release (TR). The probability of transmitter release p_{TR} is proportional to the cube of Ca^{2+}-concentration. By varying sensitivity and threshold for transmitter release probability, we modeled different fiber types [Sumner et al., 2002].

Transmitter release drives a model of presynaptic transmitter depletion. Transmitter in the vicinity of the synaptic region (the so-called readily releasable pool, RRP) is released into the synaptic cleft. The RRP is refilled by transmitter recycled from the cleft and newly produced transmitter. A large stimulus causes a transmitter release rate that is higher than the refill rate, thus depleting the RRP. Due to the depletion, the transmitter release (TR) into the synaptic cleft is reduced for the following few tens of milliseconds, an effect known as adaptation. A quantal and stochastic process models the release of transmitter into the synaptic cleft.

A single quanta of transmitter is enough to trigger a spike in the auditory nerve. The spike generation depends on transmitter release and a refractory term. An ANF can not fire twice within a shorter interval than 0.75 ms (the absolute refraction time), and after this time the probability for a spike is lowered for several tens of milliseconds (see Sumner et al., 2002).

1.4.4 Replication of Human Data

Frequency Map

The resonant frequencies of the compression resonators were adjusted to match the characteristic frequency (CF, the frequency which excites a cochlear location maximally at threshold levels) of a modified Greenwood map of the human cochlea [Greenwood, 1990]. The original Greenwood map covers frequencies between 0 Hz and 20 kHz. Ruggero and Temchin [2002] compared high-frequency cutoffs of audiograms (defined as the frequency for which the audiogram threshold was 20 dB over minimum threshold) and the highest CFs predicted by cochlear place-frequency maps for several species. They concluded that the high-frequency limit of hearing is to a large extent determined by the the highest CFs. The high-frequency cutoff of the human audiogram [Terhardt, 1979] is approximately 15 kHz. Therefore the Greenwood map was adjusted to cover a range from approximately 50 Hz to 15 kHz.

Filter Bandwidth

A common measure of frequency selectivity in psychoacoustics is the equivalent rectangular bandwidth (ERB), often expressed as a quality factor Q_{ERB} (resonant frequency divided by filter bandwidth). Recent measurements [Shera et al., 2002, Oxenham and Shera, 2003] as well as model studies [Heinz et al., 2002] have shown that suppression effects cause an underestimation of auditory filter Q_{ERB}. Suppression can be minimized by using a forward masking paradigm. (However, see Ruggero and Temchin, 2005, for a different view.) Oxenham and Shera [2003] measured Q_{ERB} at low sound levels in forward masking notched-noise experiments. For frequencies between 1 and 8 kHz the relationship

$$Q_{ERB} = 11 \left(\frac{f_{CF}}{1000 \text{ Hz}} \right)^{0.27} \tag{1.17}$$

(f_{CF} in Hz) fitted their data well. Another conclusion was that the tuning of human auditory filters appear to be much sharper than previously believed [e.g., Glasberg and Moore, 1990]. We extended Equation 1.17 to the whole frequency range modeled and adjusted $Q_{max}(x)$ of the resonators in the compression model (Equation 1.14) for the model to obtain this bandwidth. The resulting Q-values varied from $Q_{max}(x = 0) = 11.2$ to $Q_{max}(x = 35\text{mm}) = 2.38$. Q_{min} was set to 1 independent of location, resulting in a compression ranging from 84 dB for the highest CFs to 30 dB for the lowest.

The hydrodynamics model was tuned to achieve $Q_{3\mathrm{dB}}$ values increasing from 1 at the lowest CFs to 3 at high CFs according to

$$Q_{hyd}(x) = 1.94 \left(\frac{f_{res\,hyd}(x)}{1000\ \mathrm{Hz}} \right)^{0.19} \tag{1.18}$$

which results in a $Q_{10\mathrm{dB}}$ varying from approximately 0.5 to 0.9.

The correction term g_{BM} of Equation 1.8 was needed to achieve physically realistic BM displacement and match the threshold of the IHC model. In the present model, g_{BM} equals to 0.025 independent of location, which can be at least partly attributed to an overestimation of the sensitivity of the passive hydrodynamical model. Potentially, a location-dependent $g_{BM}(x)$ could be useful for adjusting thresholds of models of individual listeners, or refining models of hearing impairment.

Compression

Measurements of BM vibration in the base of the cochlea have shown very large compression at moderate and high levels [reviewed in Robles and Ruggero, 2001], with average growth rates as low as 0.2 dB/dB. Input-output functions appear to be linear at levels below 20–30 dB. For apical locations, compressive growth has been found in the chinchilla cochlea, with growth rates in the range of 0.5-0.8 dB/dB [Robles and Ruggero, 2001].

BM compression can also be estimated from psychoacoustic masked threshold experiments (for a review see Oxenham and Bacon, 2004). Several recent papers [e.g., Plack and Drga, 2003, Lopez-Poveda et al., 2003, Williams and Bacon, 2005] have estimated cochlear compression from temporal masking curves (TMC) and avoided the assumption that growth is linear for frequencies well below CF, which was generally assumed in previous work. Lopez-Poveda et al. [2003] found compression in the range between cube root and fifth root for frequencies between 500 Hz and 8 kHz. Plack and Drga [2003] concluded that there is compression of about 0.2–0.3 dB/dB at low CFs (250 Hz), and that the compression at low CFs is less frequency selective. Williams and Bacon [2005] found a compression of 0.15–0.3 dB/dB between 250 Hz and 4 kHz with little difference across frequencies. There is some controversy over whether the masked threshold experiments reflect compression of BM responses or a combination of cochlear mechanisms and compression at higher processing stages in the auditory system. Through comparisons with distortion product otoacoustic emissions, Williams and Bacon [2005] concluded that, at least for CFs above 1 kHz, their measurements reflect cochlear compression.

The compression model is constructed so that each resonator ($n = 1...4$) compresses a part of the model's total dynamic range. The values of the parameters $s_n(x)$ in the compression stage (Equation 1.12) determine the range where the model acts compressively; the relative position of the four parameters $(s_1 \ldots s_4)$, together with the maximum compression $G_{compr}(x)$, determine the slope of the growth function. The value of $s_n(x)$ corresponds approximately to the displacement that causes the Boltzmann function (Equation 1.12) to be 50% activated. The saturation parameter s_4 was set to $2.0 \cdot 10^{-8}$m independent of

location. The other saturation parameters were calculated dependent on the maximum amplification:

$$s_1(x) = \frac{s_4}{5.62 \cdot Q_{max}^3(x)} \tag{1.19}$$

s_2 and s_3 were set at uniform distances on a logarithmic scale between s_1 and s_4:

$$s_2(x) = s_4 \cdot \left(\frac{s_1}{s_4}\right)^{2/3} \tag{1.20}$$

$$s_3(x) = s_4 \cdot \left(\frac{s_1}{s_4}\right)^{1/3} \tag{1.21}$$

The resulting model exhibits between approximately third-root (low CFs) and fourth-root (high CFs) compression (compare Sec. 1.5.1 and Figure 1.7).

Hair-Cell Parameters

The fluid filter (the coupling between BM and HB motion, Equation 1.16) had a corner frequency increasing from 200 Hz at apical sites to 2 kHz at basal sites

$$\tau_{HB}(x) = \frac{1}{2\pi} \frac{1}{2000\,\text{Hz} \cdot 10^{-x/L_{BM}}} \tag{1.22}$$

The decreasing corner frequency is motivated by the stiffer IHC stereocilia found in the basal part of the cochlea [Strelioff and Flock, 1984]. The stereocilia lever gain was approximately 16 dB at the highest CFs modeled [Sumner et al., 2002], but reduced with decreasing CF according to

$$g_{HB}(x) = 2.63 \left(\frac{f_{CF}(x)}{1000\ \text{Hz}}\right)^{0.32} \tag{1.23}$$

giving no amplification at the lowest CF. The lever gain affects the threshold of the auditory nerve fibers. All other parameters were adopted from Sumner et al. [2002]. High-, medium- and low spontaneous rate fibers used the parameters of fibers H2, M2, and L2, respectively as suggested by Sumner et al. [2002].

1.5 Model Results

1.5.1 Filter Shapes

Figure 1.6 shows BM sensitivity as a function of frequency for two locations along the cochlea. Sensitivity is given as BM displacement over sound pressure level in front of the ear drum (nm/Pa). Sensitivity is largest at the characteristic frequency (CF), and varies greatly with level at CF. For a linear system all curves would be identical. The

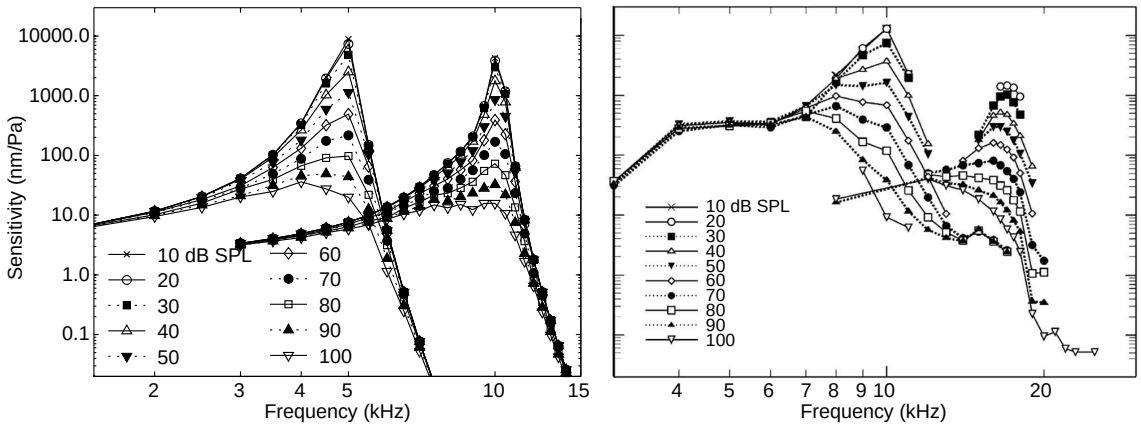

Figure 1.6: (Left) Modeled BM sensitivity as a function of frequency for BM locations 9.2 mm from the base (CF 5 kHz) and 3.2 mm from the base (CF 10 kHz). (Right) Sensitivity measurements from a chinchilla cochlea (CF 10 kHz, 3.5 mm from the base) from Ruggero et al. [1997] and guinea pig (CF 17 kHz, basal site) from Cooper and Rhode [1997]. (The figure was taken from Robles and Ruggero, 2001.) Sensitivity denotes the BM displacement relative to sound pressure in front of the eardrum. The model exhibits a boost in sensitivity at low stimulus levels and a level dependent shift in CF. Tuning the model parameters to human psychoacoustic data results in sharper frequency tuning and larger amplification in the modeled results compared to animal data.

left panel presents modeled responses for a cochlear location with CFs of 5 kHz (9.2 mm from the base) and 10 kHz (3.2 mm from the base). The right panel shows measurements in chinchilla (CF 10 kHz, Ruggero et al., 1997) and guinea pig (CF 17 kHz, Cooper and Rhode, 1997). We modeled responses approximately one octave below the measured CFs to better match the frequency range of the human model.

At high levels (100 dB), the model response is almost solely determined by the hydrodynamics model since the amplification has saturated. The travelling wave growing slowly from the cochlea base to the CF location causes the characteristic shallow low-frequency slope of the inner ear filters (approximately 6 dB/oct). After the wave reaches its maximum at its characteristic location (CL), it diminishes sharply, resulting in steep high-frequency slopes of the filters (approximately 140 dB/oct). At low levels, the compression stage is amplifying the response and sharpening the frequency selectivity. The amplification causes the filters to be almost symmetric near CF. To match human psychoacoustic data (as discussed in section 1.4.4) the model was more sharply tuned than animal data suggests, which manifests itself in more narrow-band responses at low levels. The CF shifts almost half-an-octave towards lower frequencies with increasing stimulus level, in agreement with several measurements [e.g., Ruggero et al., 1997, Cooper and Rhode, 1997]. The sensitivity of the passive response (high level responses, compression stage saturated) is comparable to the guinea pig data, but considerably lower than the chinchilla data suggests (approximately -20 dB). The sensitivity of the model is tuned by the

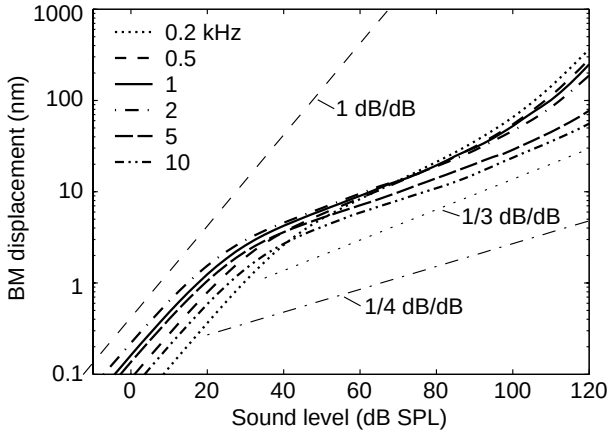

Figure 1.7: Modeled basilar membrane displacement–intensity functions for frequencies according to legend. BM displacement (rms-values) was measured at respective characteristic location. The model exhibits strong compression in the range above approximately 20 dB SPL. Compression varies from cube-root at relatively low characteristic frequencies (exemplified by 200 Hz) to fourth-root at high CFs.

correction term g_{BM} in Equation 1.8. The term was used to reduce the overall sensitivity by 32 dB ($g_{BM} = 0.025$). Narayan et al. [1998] showed that the (rate) threshold of the auditory nerve fibers in chinchilla correspond to a BM displacement in the order of 1 nm (0.26 nm for a HSR fiber and 2.7 nm for a MSR fiber). The model of synaptic process between inner-hair cell and auditory nerve has a threshold of approximately 0.5 nm. The g_{BM} term was needed to adjust the modeled BM displacement at human hearing threshold intensities to the AN threshold.

Figure 1.7 illustrates BM displacement–intensity functions for 6 different frequencies, all measured at their respective CL. At low levels (below 20–30 dB SPL) the growth is approximately linear. This is in agreement with physiological measurements in several species [e.g., Cooper and Rhode, 1992, Ruggero et al., 1997] and psychoacoustic findings [e.g., Plack and Oxenham, 1998, Plack and Drga, 2003]. At medium levels (30–80 dB), the compression varies between approximately cube-root (1/3 dB/dB) at a CF of 200 Hz to fourth-root at a CF of 10 kHz. At high levels, the growth functions become linear again as the compression stage saturates. The level at which the growth function becomes linear increases with increasing characteristic frequency. In the model this is partly because compression increases with frequency (Equation 1.15), and partly because of larger shifts in characteristic frequency with level at high CFs. (Displacement–intensity curves are measured at a fixed location along the cochlea.) From physiological measurements (reviewed in Robles and Ruggero, 2001) it seems that compression extends to at least 100 dB, but that linearization occurs at high levels.

Figure 1.8 shows auditory-nerve threshold tuning curves (TTC) for six HSR AN fibers with different CFs. To facilitate a comparison with human auditory thresholds, the simple model of the outer ear and ear canal was included for the present figure (see Section 1.4.1

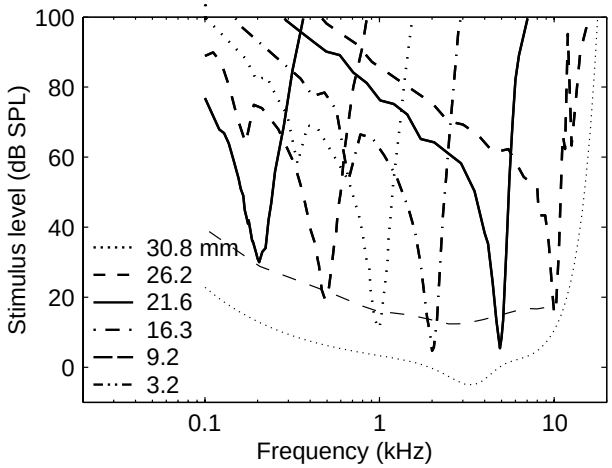

Figure 1.8: Modeled auditory nerve threshold tuning curves for five high spontaneous rate fibers with locations along the cochlea as indicated in the legend. A simple model of the outer ear and ear canal was included in the model for these calculation (see Section 1.4.1). Thin dashed line indicates the average best threshold curves for the model without the outer ear and ear canal model. The dotted line indicates human hearing threshold [Terhardt, 1979]. For calculating the threshold curves, we used a method where the threshold criteria is defined as the spontaneous rate plus one standard deviation of the spontaneous rate [Geisler et al., 1985]. Stimuli were pure tones with 250 ms duration (responses measured over the last 235 ms) and level steps were 2 dB. An "up-down" process was used, where a threshold crossing had to be repeated twice.

for details). To calculate the TTCs, we used a statistically defined criteria for determining the thresholds [Geisler et al., 1985]. The detection threshold was the level that caused a firing rate higher than the spontaneous firing rate plus one standard deviation of the spontaneous activity. The curves thus represent the auditory filters at lowest sound levels. Q-values clearly increase with increasing CF (more basal positions in the cochlea). The tail of the responses reflect the frequency response of the hydrodynamics model. The model also exhibits harmonic distortions. These are most clearly visible for the 1 kHz and 2 kHz TTCs, where the second subharmonics (333 Hz and 667 Hz respectively) elicit stronger responses than adjacent frequencies.

The thin dotted line indicates hearing threshold of humans [Terhardt, 1979]. The thin dashed line indicates the averaged best-threshold curve (BTC, Liberman, 1978), the most sensitive threshold tuning curves, of the model *without* outer ear and ear canal (i.e., the way the model is used in the rest of this chapter).

The transfer function describing the ratio of sound pressure in front of the ear drum and sound pressure of a plane wave in a free field has a maximum amplification of more than 15 dB [Shaw, 1974, Mehrgardt and Mellert, 1977, see also Section 1.4.1]. The strongest amplification occurs for frequencies between 2 kHz and 5 kHz. The simple model used here does not quite reach that amplification. The increase in model thresholds at low

frequencies results partly from middle ear filtering and partly from the reduced coupling gain between BM and the inner-hair cell stereocilia (compare Sec. 1.4.4). It should be noted that the threshold curves in Figure 1.8 are *rate* threshold curves. At low frequencies, where phase-locking is prominent, thresholds are likely to be 10–20 dB lower than a rate-threshold indicates [Johnson, 1980]. In summary, the model was designed to approximately match hearing thresholds when including a model of the external ear and considering phase-locking effects.

1.6 Structure of the Thesis

The rest of the thesis is organized as following: In Chapter 3 we modeled the first neuronal processing stage in the Ventral Cochlear Nucleus: the Onset Neurons. Onset Neurons receive input from Auditory Never Fibers (ANFs) and forward the information to higher processing stages of the brain using discrete spike trains. Chapter 2 described a method for improving the original model using a realistic offset adaptation, which is essential for the proper functioning of the ONs. Improvements in the model were compared to experimental data whenever possible. In Chapter 4, we harnessed the tool of automatic speech recognition to qualitatively measure the information processing in the modeled inner ear model. We used features extracted from different stages of the model and different speech recognition engine for the speech recognition tasks. We showed how realistic auditory modeling and the proper acoustical model could help to improve speech recognition performance. Results were also compared with human performance to show the performance gap. In Chapter 5, we used information theory to quantitatively evaluate the information processing using our auditory model. Information theory estimates information carried by the spike trains of ONs without making assumptions. It helped to explain at least partially why there is still a big performance gap between automatic speech recognition and human performance. Chapter 6 was an extension of the study in Chapter 5, where we applied information theory on different types of neurons and analyzed their properties from the perspective of information transmission.

2 Offset Adaptation

Abstract[1)]

Recent pool models of the inner hair cell synapse fail to reproduce a silent period after an intense stimulus. This has important consequences in the next processing step, the cochlear nucleus which receives direct input from auditory nerve fibers (ANFs): Onset Neurons in the ventral cochlear nucleus (VCN), modeled with a detailed Hodkin-Huxley model Rothman and Manis [2003a] do not reliably respond to amplitude modulated signals in the frequency region above 4 kHz. An analytical method to achieve the desired adaptation properties of the auditory nerve was implemented into the existing model of the peripheral auditory system. The revised model produces more realistic offset adaptation in accordance with physiological measurements, while generating the same onset adaptation. Analysis on the model output shows that the auditory nerve fibers of the enhanced model fire more synchronously to the peak of the stimulus. The analysis of the synchronization index showed that the model output was able to replicate the best results from physiological measurements. Such output meets the criterion of preferable input to onset neurons and leads to a considerable modulation gain (up to 40 dB) for these cells. With this extension enabled onset neurons to extract amplitude modulations in the high frequency range, providing important information about high formants in speech signals. In conclusion, offset adaptation is an important feature of ANF responses which is essential for neuronal processing of auditory signals by onset neurons in the VCN.

[1)] A modified version of this chapter has been published in Interspeech 2008 [Wang et al., 2008].

2.1 Introduction

Onset Neurons (ONs) located in the first neural processing stage after the inner ear and are known for their distinct temporal processing capabilities. They receive neural stimulations from the auditory nerve fibers and relay the information to further processing stages in the brain (see Figure 2.1). The onset neurons in our system used a single compartmental model including four major Hodgkin-Huxley type icon channels, with parameters taken from Rothman and Mannis [Rothman and Manis, 2003a]. We corrected conductances and time constants to a body temperature of 38° and solved the differential equations in the time domain. Analysis of ONs using reverse correlation techniques show that ONs respond preferably to stimulus onset while cease to fire at stimulus with a constant intensity. In natural speech processing, ONs enhances the periodicity of voiced speech by very reliably extracting the pitch frequency of the speaker. In Chapter 3 we will describe the onset neurons in great detail.

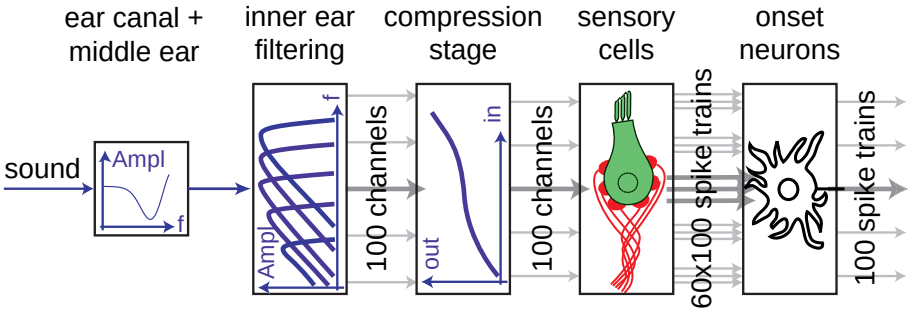

Figure 2.1: Schematics of the auditory model. The model exhibits 100 frequency channels, which are coded by multiple ANFs. A single VCN neuron is excited by ANFs originating from a distinct frequency channel.

One of the most critical processing steps during auditory sound processing occurs at the inner hair cell (IHC) synapses: here the mechanically pre-filtered analog sound signal is converted into discrete nerve action potentials which propagate along the auditory nerve fibers (ANF) to the brain. This conversion induces massive information loss – or to phrase it positively – information reduction. As any information lost during this process is no longer available for neuronal processing, it is important to understand and model the underlying principles correctly.

Physiological studies have shown that the auditory nerve response to a constant intensity stimulus is typified by very strong firing at the stimulus onset, and a very rapid decline shortly after the onset, which then slows down over several tens of milliseconds. After the stimulus offset, auditory nerve fibers remain silent for a short time and then recovers slowly to its spontaneous activity (Figure 2.2).

However, the IHC-AN model from Meddis previously used in our peripheral auditory system failed to reproduce offset adaptation. Therefore, phase-locking to the input stimulus and to amplitude modulations was not sufficient. As a result, ONs in the cochlear nucleus hardly fired at frequencies above about 2 kHz.

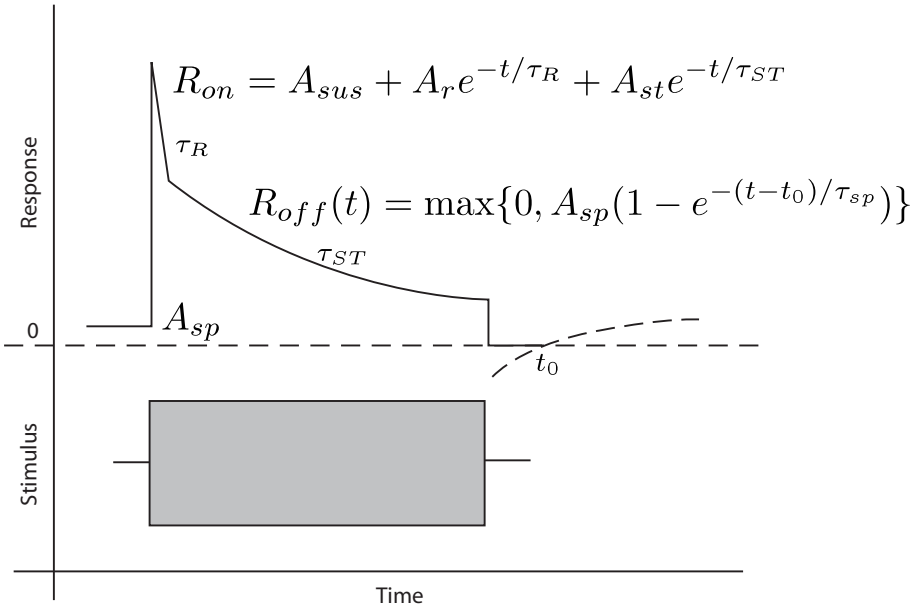

Figure 2.2: Schematic diagram of the response of AN fiber to a tone burst [Harris and Dallos, 1979].

2.2 Modeling IHC-AN Synaptic Adaptation

The characteristics of IHC-AN synaptic adaptation depend on stimulus intensity, duration and previous stimulation history. Adaptation observed at the onset of ANF responses to a tone burst is explained by the depletion of synaptic vesicles in a pool close to the release site (immediate store, Figure 2.3) in the afferent synapse. The observed time course in the firing rate of ANF responses to tone bursts can be characterized by two exponential components [Westerman and Smith, 1984] (see Figure 2.2):

$$R_{on} = A_{sus} + A_r e^{-t/\tau_R} + A_{st} e^{-t/\tau_{ST}} \tag{2.1}$$

where A_r and A_{st} are the two exponential components of rapid and short term adaptation, τ_R and τ_{ST} are the respective decay time constants, and A_{sus} is a steady-state component.

Offset adaptation can also be described with an exponential recovery component with different time constant than the onset, after a "dead-time" period:

$$R_{off}(t) = \begin{cases} 0; & t < 0 \\ A_{sp}(1 - e^{-(t-t_0)/\tau_{sp}}); & t \geq t_0 \end{cases} \tag{2.2}$$

where A_{sp} is the spontaneous rate, t_0 is the dead-time period, and t_{sp} is the recovery time constant of the offset adaptation.

Given the diversity of auditory nerve fibers with different adaptation properties and the

complexity of the synaptic adaptation, it has been a challenging task to model the synaptic dynamics successfully. In the peripheral auditory system that we showed in Section 1.4, we implemented a pool model as proposed by Meddis [Meddis, 1986, 1988], which very successfully describes the adaptation characteristics of auditory nerve fibers to some extent.

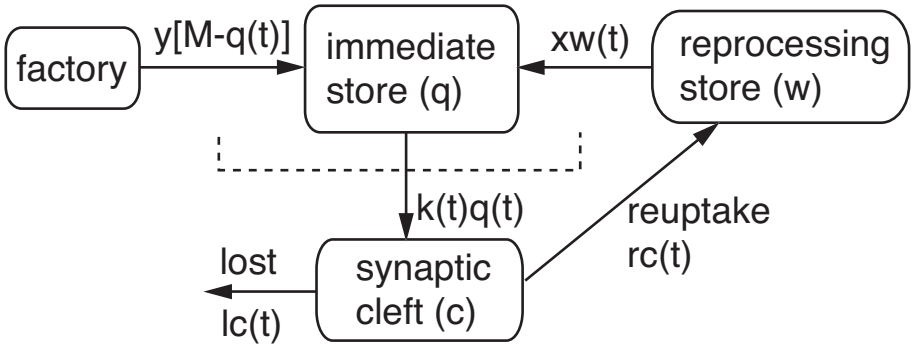

Figure 2.3: Schematics of the IHC-AN model. The immediate store (q) (maximum size: M) is refilled both by the global pool at a rate of $y[M - q(t)]$ and by the reprocessing store (w) at a rate of $xw(t)$. The transmitter in the synaptic cleft (c) is either lost at a rate of $lc(t)$ or recycled at a rate of $rc(t)$ into the reprocessing store.

The model proposed by Meddis has three neurotransmitter reservoirs which are specified by the synaptic transmitters in the reservoir and arranged in a circle: the immediate store (q), the synaptic cleft (c), and the reprocessing store (w)(see Figure 2.3). The model output is proportional to the rate of transmitter release from the immediate store (q) to the synaptic cleft (c), given by $k(t)q(t)$, where $k(t)$ is the only stimulus dependent variable in the Meddis model. $k(t)$ describes the fusion rate of synaptic vesicles (mediated by Ca^{2+}-influx into the cell) which is usually specified as a function of intracellular inner hair cell voltage [Sumner et al., 2002].

The immediate store q (maximum size: M) is refilled both by the global pool at a rate of $y[M - q(t)]$ and by the reprocessing store (w) at a rate of $xw(t)$. The transmitter in the synaptic cleft is either lost at a rate of $lc(t)$ or recycled at a rate of $rc(t)$ into the reprocessing store.

These replenishment and release of the neurotransmitter can be described with the following equations:

$$\frac{dq}{dt} = y(M - q(t)) + xw(t) - k(t)q(t) \tag{2.3}$$

$$\frac{dc}{dt} = k(t)q(t) - (l + r)c(t) \tag{2.4}$$

$$\frac{dw}{dt} = rc(t) - xw(t) \tag{2.5}$$

These equations can be transformed into the Laplace domain (Equation 2.6) and solved analytically, with a high-frequency tone burst as the stimulus input, wherein the IHC voltage is assumed to be constant after the onset (denoted as k_2). For more details please refer to the paper from Zhang and Carney [2005].

$$Q(s) = \frac{(sq(0^-) + yM)(s+x)(s+l+r) + c(0^-)rxs + w(0^-)xs(s+l+r)}{s(s+x)(s+y+k_2)(s+l+r) - k_2rxs} \tag{2.6}$$

the resulting characteristic function of $q(t)$ can be represented by

$$q(t) = \Phi_0 + \Phi_1 e^{-t/\tau_1} + \Phi_2 e^{-t/\tau_2} + \Phi_3 e^{-t/\tau_3} \tag{2.7}$$

where $-1/\tau_i$ are poles of $Q(s)$. The values τ_i and Φ_i can be calculated from $Q(s)$ directly.

Equations 2.6 and 2.7 can be further simplified and the following result was given by Zhang et al.

$$q(t) = \Phi'_0 + \Phi'_1 e^{-t/\tau'_1} + \Phi'_2 e^{-t/\tau'_2} \tag{2.8}$$

The simplified equation generates indistinguishable results as long as $(l+r) > 5000$, which can be always guaranteed as $l + r$ used in the Meddis model were usually greater than 15000. The simplified equation provides an accurate description of the Meddis model. Equation 2.8 has two exponential components with different time constants which are the same as in Equation 2.1.

Therefore, the response of the model to a stimulus with constant intensity can be determined analytically. Then, the relationship between model parameters and adaptation characteristics can be established. For a stimulus with constant intensity, these two groups of parameters can be derived from each other in a mutual fashion. Thus any desired adaptation responses can be achieved using appropriate parameters, and vice versa.

An underlying assumption of the model is that all model parameters are constant, except the stimulus dependent permeability $k(t)$. The model parameters are derived from adaptation parameters at a given level (a medium or high level for high-spontaneous-rate fibers and a very high level for low-spontaneous-rate fibers). The adaptation characteristics then change with the stimulus-dependent parameter $k(t)$ accordingly.

There has been less attention paid to the modeling of offset adaptation at the IHC-AN synapses. As there is no further data available which characterizes the dynamics of offset adaptation, we choosed to apply identical parameters, e.g. the same characteristic function for both onset and offset adaptation.

Physiological studies suggest that AN fibers with medium- or high-spontaneous rate usually stop firing immediately after the offset, and recover slowly after the dead-time period with a time constant longer than the short time adaptation. However, the offset adaptation property of Meddis model is limited by the model structure. Since $q(t)$ can not

be negative in Equation 2.8, the amplitudes of Φ'_1 and Φ'_2 are limited by Φ'_0. The rapid component of the model recovery function causes the synapses to recover very quickly after stimulus offset.

In order to achieve a more physiologically consistent offset adaptation, a shift value A_{shift} was added to allow for negative rate at signal offsets. In a next step negative outputs are set to 0, which represents the dead-time period. Therefore, the characteristic function of the output becomes:

$$R(t) = max[k(t)q(t) - A_{shift}, 0]; \tag{2.9}$$

In order to keep the onset adaptation unchanged, the synaptic output becomes:

$$\begin{aligned} k(t)q(t) &= R_{on} + A_{shift} \\ &= A_{shift} + A_{sus} + A_r e^{-t/\tau_R} + A_{st} e^{-t/\tau_{ST}} \end{aligned} \tag{2.10}$$

We then re-calculated parameters of the model for the new adaptation parameters and subtracted A_{shift} from $k(t)q(t)$ to get the output of the pool model, as in Equation 2.9.

Thus, by including this shift, the same equation can be used to represent the onset and offset adaptation in the model. In summary, the new model achieved the desired offset adaptation while keeping the onset adaptation untouched.

2.3 Results

The modified IHC-AN synaptic model produces the same onset adaptation but with a physiologically more realistic offset adaptation (see Figure 2.4). The time constants of rapid adaptation and short term adaptation are 1 ms and 54.7 ms, respectively. Figure 2.4 compares the traditional and enhanced model of adaptation with physiological data. Traditional adaptation models only show a depression of spontaneous activity and an exponential recovery after a tone burst. For the enhanced adaptation model, ANF responses were silenced during the dead-time period after signal offset and then slowly recovered to spontaneous activity, in accordance with physiological experiments [Kiang et al., 1965] [2)].

An important property of the output spike train from the ANFs is the synchronizing or phase-locking of the spikes to sinusoidal waves, defined by the synchronization index [Goldberg and Brown, 1969]. We recorded the phase relation between the stimulus and the output spike trains. Then each spike can then be considered as a vector of unit length with a phase angle θ_i, $0 \leq \theta_i \leq \pi$. In this case, the n vectors characterizing a

[2)] Note that the enhanced adaptation model adds only one parameter, A_{shift}. The model is tuned to keep the spontaneous rate of the fibre constant; the driven rate changes slightly.

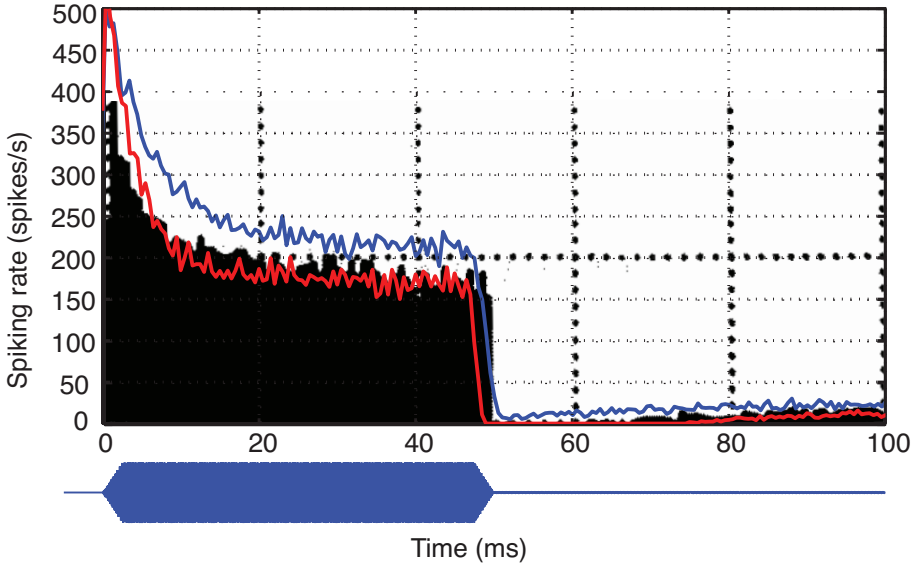

Figure 2.4: Response of an auditory nerve fiber (black area, measurements from cat [Kiang et al., 1965]) to a tone burst at CF (10.34 kHz, 39 dB) and model output with enhanced offset adaptation (red line). Note that after the tone burst the spontaneous activity is silenced for about 15 ms and recovers slowly thereafter, which is not predicted by the adaptation model of Sumner et al. [Sumner et al., 2002] (blue line) and requires the introduction of an effect termed "offset adaptation". In our calculations we introduce offset adaptation by adding a threshold to the pool model of adaptation, which generates a realistic "offset adaptation".

spike train can be treated as a distribution on a unit circle. We then calculated the mean vector. The direction of the mean vector is a measure of mean phase relation between the stimulus and the output spike trains. The length of the mean vector,

$$R = \frac{\sqrt{(\sum \cos\theta_i)^2 + (\sum \sin\theta_i)^2}}{n} \tag{2.11}$$

provides a measure of synchronization or phase locking. The parameter R, called the "synchronization index" or the "vector strength" takes value from 0 to 1. A value of 1 implies perfect phase synchronization of the output spikes; a value of 0, given certain restrictions, implies spikes occur randomly throughout the stimulus.

An extensive analysis of the synchronization index revealed that the modified auditory nerve fiber model more precisely phase lock to the input stimuli. Both models with and without the enhanced offset adaptation achieve high synchronization indices in the low frequency region ($\leq 1\,kHz$) with values in the range of measurements [Johnson, 1980]. In the original model, the synchronization index degrades drastically in the frequency range above 1 kHz and lies far below experimental data (see Figure 2.5). The model with enhanced offset adaptation greatly improved the synchronization and the model data tightly followed the best results achieved from measurement also at high frequency region.

Over the whole hearing range, the modified model achived a higher synchronization index than the original one.

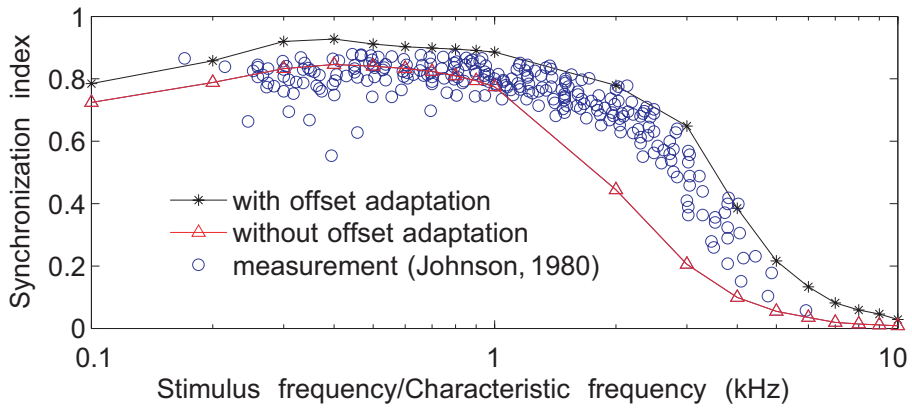

Figure 2.5: Synchronization index of auditory nerve action potentials. Input stimuli are pure tones at different frequencies. Synchronization indexes were calculated from output spike trains of neurons whose characteristic frequencies corresponded to the input stimuli. Modeled results are shown with and without adaptation, together with measurements from Johnson [Johnson, 1980].

The enhanced phase-locking property of auditory nerves is vital for further neural processing stages. We connected the octopus neurons, modeled with a detailed Hodgkin-Huxley model [Rothman and Manis, 2003a] (for a detailed description see Chapter 3), to the ANFs, and found that they responded in the frequency region above 3 kHz only when offset adaptaion was included in the IHC-AN model (compare panels c and d in Figure 2.6). Octopus neurons require a quiet period of about 1–2 ms before they fire [Hemmert et al., 2005], an effect for which offset adaptation is essential.

The offset adaptation in the modified model accounts for enhanced phase locking to the stimulus envelope in AN fibers. AN fibers become less responsive during the dip of the envelope and therefore more synchronized to the peak of the amplitude modulation. As speech stimuli can be essentially regarded as amplitude modulated signals, output of AN fibers become more synchronized at the onset of the speech signal and silenced at the offset, which then provides the preferable input for onset neurons, especially the so called bushy cells and the octopus cells.

Figure 2.6 shows the outputs from auditory nerve fibers and octopus cells for both models. Octopus cells exhibit band-pass characteristics and react preferably to signal onset, while suppressing steady-state activities [Hemmert et al., 2005]. In the original model (left column), the auditory nerve spike trains are always active, also in stimulus dips, which almost completely suppressed responses from octopus neurons. With offset adaptation, octopus cells respond reliably even in high frequency regions, extracting the important formant information and coding temporal information of the speech signal.

We calculated the modulation gain [Joris and Yin, 1992] of octopus cells to study the effect

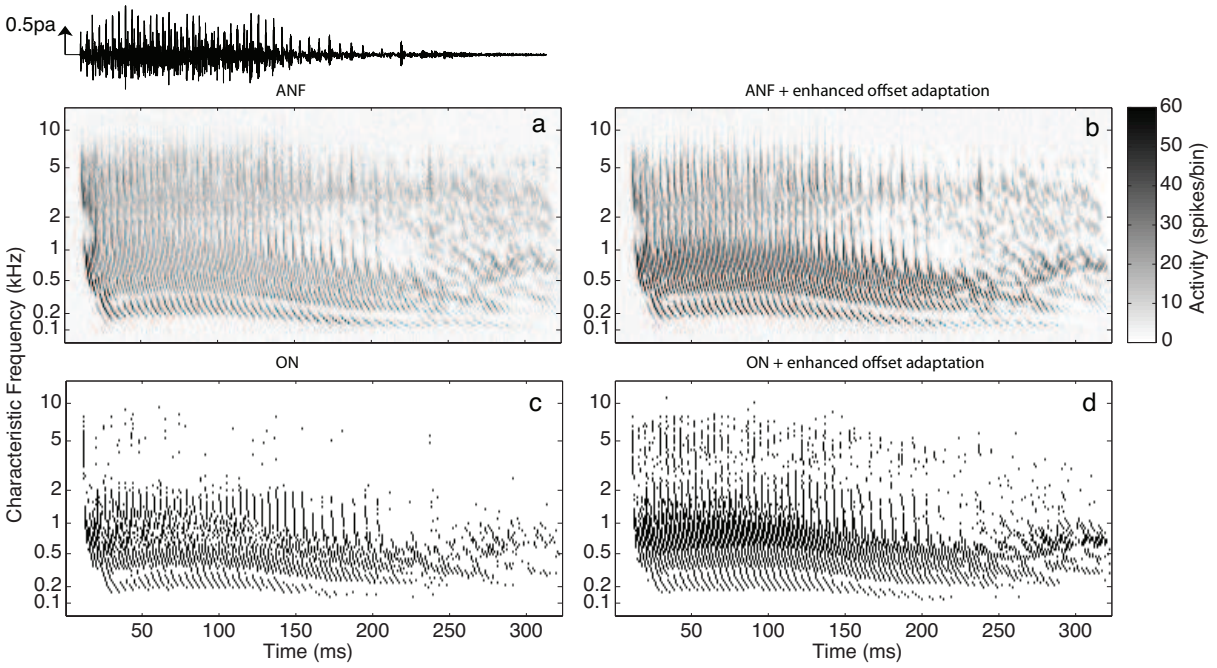

Figure 2.6: Responses from ANFs (upper row) and ONs (lower row) for our model of auditory sound processing with (right column) and without (left column) offset adaptation. For each frequency channel, we plotted responses of 60 ANFs and one ON innervated by them. The stimulus was an "a" from a female speaker (ISOLET). The number of spikes falling in 1 ms time bins was represented in gray scale for the ANF response (see color bar top right).

of offset adaptation on onset neurons. Modulation gain was defined as the following,

$$Modulation\,gain = 20\log[(modulation\,of\,response)/(modulation\,of\,stimulus)] \quad (2.12)$$

Usually the modulation of the response is measured by the synchronization index R, while the modulation of the stimulus is given by the modulation depth m of amplitude modulated sinusoidal signal. However, since R of a half-wave-rectified AM waveform with $m = 1$ is 0.5, modulation depth and R are not equivalent metrics. To obtain a gain of 0 dB when the modulation of the response equals to that of the stimulus, a factor of 2 must be included, i.e.,modulation gain $= 20\log(2R/m)$) [Rees and Palmer, 1989]. A disadvantage of this measurement is that the synchronization index may fail to reveal the real phase-locking capability in some cases. For example, when neurons spike only once, the vector strength in Equation 2.11 would equal to 1 and the output would be regarded as perfectly synchronized, while in fact the neurons almost totally lose the phase locking ability to the given stimulus. Therefore, when calculating the synchronization index (mean vector strength), we divided the strength of the vector by the number of modulated periods, instead of by the real number of spikes (n as in Equation 2.11), as we expect one spike per onset if the neurons phase lock to the modulation perfectly. For ANFs we don't have such a problem since ANFs always generate large numbers of output spikes.

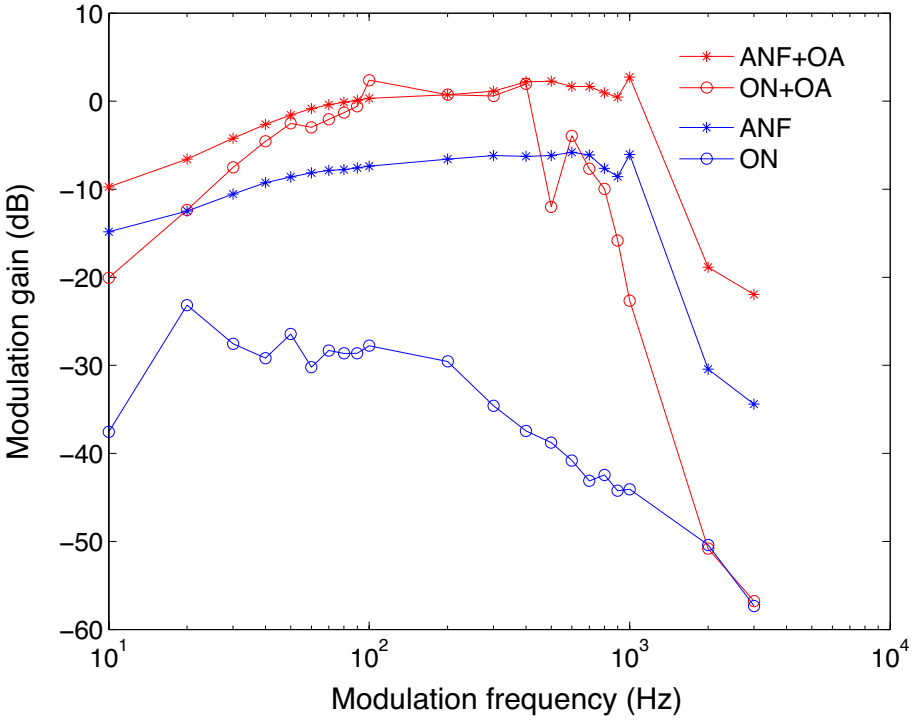

Figure 2.7: Modulation gain for AN fibers and octopus neurons. Carrier frequency is 10 kHz. So are the characteristic frequencies of the auditory nerve fiber and octopus neuron. The sinusoidal carrier signal is modulated with different modulation frequencies, with modulation depth equal to 1. The intensity of the input signal is kept constant at 70 dB(A).

The modulation gain is shown in Figure 2.7. The enhanced offset adaptation leads to a moderate 5-10 dB modulation gain for auditory nerve fibers, whereas for octopus neurons there is a very impressive improvement in terms of modulation gain. The extra gain brought by offset adaptation amounts to as much as 30–40 dB in the frequency range between 100 Hz and 1 kHz. Octopus neurons with the enhanced offset adaptation model show the highest modulation gain in the medium frequency region. The filter described by the modulation gain curve resembles the filter shape that we got from the frequency response of octopus neuron's reverse correlation function (compare Figure 3.6).

We also calculated sychronization indexes for the octopus neurons. However, octopus neurons generally fired only on signal onsets at middle to high CF when using pure tone stimuli, hence no sychronization indexes can be calculated at those CFs. On the other hand, modulation gains provide a good measurement for octopus neurons as they still respond regularly to AM signals. To characterize the properties of octopus neurons, we used AM signals with different carrier frequencies but the same amplitude modulation (modulation depth: 1, modulation frequency: 100 Hz). In this case the signals stimulate the octpus neurons across a wide range of characteristic frequencies.

Octopus neurons exhibit very different response properties below and above about 1 kHz.

Below about 1 kHz they are fast enough to phase lock to the carrier frequency of the input stimulus. Above 1 kHz, octopus neurons only fire at the onset of a sound stimulus. In addition, for amplitude modulated stimuli octopus neurons phase lock to amplitude modulations.

In a next step, we compared synchronizatin indexes for both octopus neurons and ANFs. Since octopus neurons do not respond to pure tone signals at middle to high CFs, different signals were used to measure the phase locking properties. At low frequencies, where octopus neurons lock to the carrier signal, we used pure tone stimuli as input and measured synchronization relative to the carrier signal; at middle to high frequencies, where octopus neurons respond well to AM signals but not to pure tone stimuli, synchronization indexes were measured relative to the modulated signal.

Figure 2.8 shows synchronization indexes for ANFs and octopus neurons. At characteristic frequencies between 100 Hz and 1 kHz, outputs of the octopus neurons are highly synchronized to the carrier signal. Octopus neurons enhance phase locking to sinusoidal signals in comparison to ANFs. At very low frequencies, from 100 Hz to 500 Hz, octopus neurons are able to fire in every stimulus cycle with extremely high phase synchronization. As the carrier frequency increases, synchronization indexes of both octopus neurons and ANFs drop.

The dashed lines show synchronization indexes to AM signals. Again the octopus neurons were able to achieve very high synchronization index values at high frequencies above 1.1 kHz. They phase lock very precisely to the AM signal, and fire almost each cycle of the modulation. The high synchronization values to AM signals show that the octopus neurons enhance the amplitude modulation. They relay precise temporal information extracted from auditory nerve fibers with much less spikes to high processing stage.

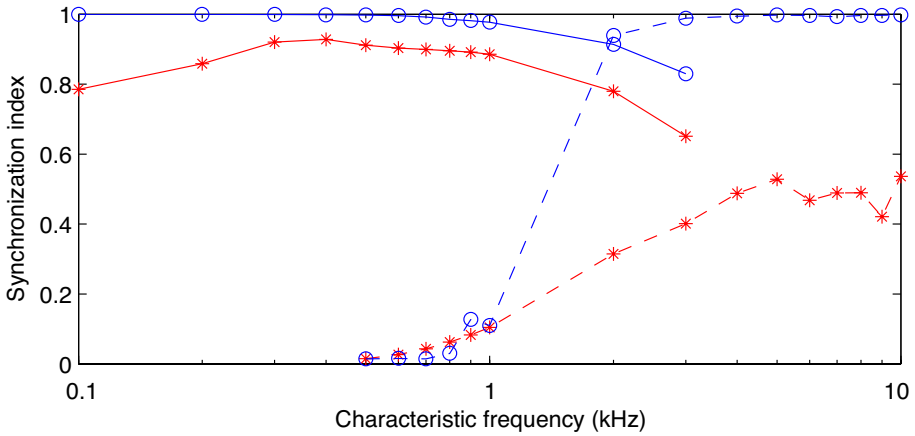

Figure 2.8: Synchronization indexes for octopus neurons (circles) and ANF (crosses). Solid lines show synchronization to the carrier frequency using pure tone as stimuli; while dashed lines show synchronazation relative to the amplitude modulation using AM signal (full modulation, 100 Hz modulation frequency) as input.

2.4 Discussion

Physiological studies suggest that AN fibers with medium- or high-spontaneous rate usually stop responding immediately after offset and recover slowly after the dead-time period with a time constant longer than that of short term adaptation. The model originaly used in our peripheral hearing system, however, failed to reproduce this behaviour. The rapid component of the model recovery function cause the synapses to recover very quickly after a tone burst is switched off. This behaviour, at the next neuronal processing stage, fails to eleicit responses of ONs (bushy cells and octopus cells) to amplitude modulated signals like speech.

In this chapter, we extended the model of the peripheral hearing system, with a better offset adaptation proposed by Zhang et al [Zhang and Carney, 2005]. By introducing a shift to the original IHC-AN model proposed by Meddis, we were able to use the same equations to represent both onset and offset adaptation. The modified model enhances the offset adaptation while generating the same onset adaptation.

The modification of the Meddis model replicates a dead-time period of several tens of milliseconds after turning off a tone burst in the auditory nerve response, according to physiological measurements.

We calculated the synchronization indexes of the auditory nerve spikes from the original model and the enhanced model. The enhanced model shows an obvious improvement in synchronization of the output spike trains to pure tone stimuli. The output spikes of the enhanced model show better phase locking to pure tone stimuli. The improvement in synchronization index is especially large in the high frequency range from 1 kHz to 5 kHz, where the phase locking ability of the original model drops dramatically. This improvement proves to be vital to the next neural processing stage to provide the appropriate stimulus for onset neurons in the cochlear nucleus.

The relatively modest improvement on the auditory nerve synapse model leads to considerably more realistic responses of the onset neurons, especially for bushy and octopus neurons. This is because these neurons fire preferably to the well synchronized signal onset in an all-or-none manner. They work as synchrony detectors rather than just integrating the input from auditory nerve fibers. The enhanced model generates more synchronized ANF spikes as indicated by the synchronization ANFs fire more synchronously during the onset of the signal while they fire less in the dip, which is a very impotent property for ONs to work at the middle to high frequency range. The benefit of incorporating offset adaptation is obvious: bushy and octopus cells failed to fire in the frequency region above about 1 kHz in the original model, while with enhanced offset adaptation, they nicely extract synchronies.

The modulation gain shows considerable improvement in the performance of the ANF and the octopus neuron. The improvement for the octopus neurons is the most impressive: the highest gain in modulation is as much as 40 dB. Besides, the enhanced model leads to a more realistic modulation transfer function of octopus neurons, which resembles the

filter shape of octopus neurons calculated using reverse correlation techniques.

In summary, offset adaptation is an essential feature found in auditory nerve responses. It improves phase locking and the coding of amplitude modulations in the auditory nerve and, moreover, is essential to drive onset neurons in the next processing stage, the cochlear nucleus.

3 Modeling the Onset Neurons

Abstract

This chapter presents a model of the cochlear nucleus onset neurons with Hodgkin-Huxley type ion-channels with the major ionic conductances characterized by Rothman and Manis. In particular we studied the octopus neurons which are famous for their distinctive temporal processing property. Octopus neurons reject steady-state excitation and fire on signal onsets with extremely high temporal precision. They phase lock to pure tone stimuli with high reliability at frequencies below 1 kHz. They were also able to react to amplitude modulated signals with entrainment up to 1 kHz. When given vowels as an input, octopus neurons fire very reliably at almost each glottis stroke, extracting faithfully the pitch frequency of the speaker. We applied reverse-correlation technique to analyze the octopus neurons, and found that the modeled octopus neurons perform a band-pass filtering on the incoming signal, with a low frequency slope of approximately 6 dB/octave. This indicates that in general the octopus neurons process the first derivative of the input signals.

3.1 Introduction

Neurons are responsible for the signal processing in the mammalian nerve system. Action potentials from multiple ANFs reach the synapses of ONs. At the end of the axon of a neuron there are many synaptic terminals, connected to the dendrites of the other neurons.

The synaptic terminals and the dendrites are separated by plasma membranes. The voltage difference at the membrane is known as the membrane potential V_m. The neuron's resting potential is between -40 mV and -90 mV. The majority of nerve cells generate a series of brief voltage pulses also referred to as action potentials or spikes. They are all common in their all-or-none depolarization of the membrane beyond the threshold; we assume the threshold to be $T_h = 0$ V. If the voltage exceeds the threshold, the membrane executes a voltage trajectory. The amplitude of the action potential is almost independent of stimulus intensity. All the information from one neuron to the other is transferred by the spikes. The firing of the action potential is binary and follows the all-or-none law. The depolarization time of the membrane potential is rather short. Nevertheless it is impossible to evoke another action potential immediately after firing one, due to the refractory period.

The activation of the action potentials are mainly governed by the voltage-gated and strongly nonlinear ion-channels. The initial stimulus opens calcium channels, Ca^{2+} enters the presynaptic part and activates enzymes, which cause the vesicles close to the synaptic cleft to release their neurotransmitter. They diffuse to the postsynaptic membrane and bind to postsynaptic membrane receptors. At excitation, the activated receptors cause an activation or even an action potential in the postsynaptic neuron. The electrical behavior of octopus neuron is dominated by two voltage-sensitive conductances [Golding et al., 1999]: A low-threshold potassium channel (K_{LT}) with activation kinetics in the order of 2 ms [Bal and Oertel, 2001, see also Svirskis et al., 2004], and a hyperpolarization-activated, mixed-cation channel (I_h, Bal and Oertel, 2000). Both channels are already activated at rest, but react to voltage changes in different directions. The activation of both channels at rest gives the cell an unusual low input resistance. When the ON membrane is depolarized, they elicit an initial action potential, but thereafter K_{LT} compensates input currents and keeps the membrane potential below spiking threshold. For octopus neurons, action potentials of the auditory nerve elicit only extremely brief activation of postsynaptic currents [compare Rothman and Manis, 2003a]. The octopus neuron model presented here (and used in automatic speech recognition experiments in Chap. 4) relies only on ionic conductances which were fully characterized in physiological experiments and requires no further hypothetic mechanisms. It is a single-compartment model, and relies on Hodgkin-Huxley-type ion channels. The conductances and channel dynamics are based on measurements of VCN neuron conductances [Bal and Oertel, 2000, 2001, Rothman and Manis, 2003a]. Figure 3.1 shows the equivalent circuit diagram for modeling the ONs.

ONs extract the spectral-, temporal-, and spatial information coded by the AN spike trains

and transmit these information to higher processing stages. ONs locate in the cochlear nucleus (CN), the first neuronal processing stage after the inner ear. The dendrites of ONs tap the tonotopic array of auditory nerve fibers systematically, maintaining a tonotopic map of all frequency range [Oertel et al., 2000]. There are a large number of specialized cell types in CN, which are generally spatially segregated. Different cell types are associated with various organization and specification of the synaptic terminals. Auditory nerve terminals that innervate bushy and stellate cells are variable in size and shape. In the octopus cell area, in contrast, terminals of auditory nerve fibers are uniformly small boutons. In this and the following chapter we mainly deal with octopus cells. But the modeling method also applies to other different ONs. It has been found that in mice about 200 octopus cells sample the array of about 12,000 auditory nerve fibers [Willott and Bross, 1990, Ehret, 1979]. As all auditory nerve fibers have been observed to terminate in the octopus cell area, octopus cells receive on average at least 60 inputs [Lorente de No, 1933, Brown and Ledwith, 1990][1)]. The convergent input from a relatively large number of auditory nerve fibers via small terminals is reflected in the response of octopus cells. Recordings from octopus cells indicate that they receive subthreshold synaptic input, therefore inputs from multiple auditory nerve fibers had to sum to produce an action potential in octopus cells. Octopus neurons have extraordinarily fast membrane time constant (0.2–0.3 ms) and react to the simultaneous firing of multiple auditory nerve fibers very fast, acting as coincidence detectors and greatly enhancing the precision of timing relative to a single ANF [Golding et al., 1995].

Octopus neurons have some distinctive characteristics. Unlike some other neurons (e.g., bushy neurons and chopper neurons) which exhibit sustained activity to continuous excitation, octopus neurons show strong onset responses: they fire only once at the onset of high frequency pure tones. For the remainder of the tone burst they fire with very low rate (less than 10 spikes/s) or do not fire at all [e.g., Godfrey et al., 1975, Rhode and Smith, 1986]. When given low frequency pure tones as stimuli, octopus neurons fire with entrainment (see Chapter 2) [2)]. Entrainment can be sustained up to 800 kHz (see Chapter 2 and [Rhode and Smith, 1986]), resulting in a very high firing rate of about 800 spikes/s. In comparison, the auditory nerve saturates at about 300 spikes/s. Octopus neurons also show remarkably exact and reliable responses to amplitude modulated signal [Rhode, 1994] and speech like stimuli [Rhode, 1998].

The distinctive temporal processing properties of octopus neurons make it very interesting to us; they suppress spontaneous and sustained activity while firing preferably at onset of input stimuli, a feature which could be essential for sound processing in noise. Their detection of signal onsets is important for sound localization as well as for the perception of speech and music. Identical onset times have been proposed as one of the key cues in auditory scene analysis [Bregman, 1990, Cooke and Ellis, 2001]. Another important cue in auditory scene analysis is periodicity. The octopus neurons distinctive response to

[1)] The number of auditory nerve inputs onto octopus cells may as well be several times 60 because many auditory nerve fibers probably innervate multiple octopus cells.

[2)] Entrainment means that they not only phase lock to the input stimuli, but also respond to each cycle of the stimulus with exactly one spike.

periodic stimuli suggests that they play an important role in periodicity extraction. In fact, octopus neurons phase lock to the pitch frequency of the speakers with high precision and reliability.

3.2 Modeling

In this section we will provide the basic mechanism in an onset neuron. We will also show how we modeled the neurons in a mathematic way. Detailed physiological description of the neurons and how the ion channels work is not included here. Readers are referred to Klinke and Silbernagl [1994] for a fuller explanation.

We modeled the octopus neurons using a single compartment model with Hodgkin-Huxley type ion channels. We used parameters of the conductances of the neuronal channels measured by Rothman and Manis [2003a,b,c]. The steady-state and dynamic equations were also based on their studies. The octopus neuron was named as Type II-o model in their study. In principle it is a hybrid of the generic VCN model with a very large low threshold potassium conductance, coupled with an implementation of the hyperpolarization-activated mixed-cation conductance measured by Bal and Oertel [2000].

3.2.1 Hodgkin-Huxley Model

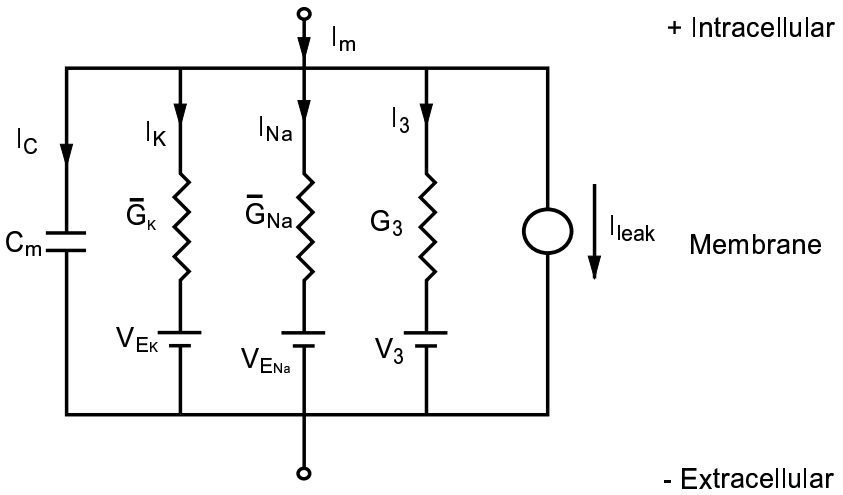

Figure 3.1: **Equivalent circuit diagram**

The most notable work to describe the ionic mechanisms underlying the initiation and propagation of action potentials was done by Hodgkin and Huxley in Cambridge, England in 1952. The Hodgkin-Huxley model is realized in a equivalent circuit in Figure 3.1 ($I_3 = G_3 = V_3 = 0$). In the **Hodgkin-Huxley-Model** the total membrane current $I_m(t)$

is the sum of the ionic current $I_{ionic}(t)$ and the capacitive current I_C,

$$I_m(t) = I_{ionic}(t) + C_m \frac{dV(t)}{dt} \tag{3.1}$$

According to their experiments with a squid giant axon, they posted a phenomenological model Koch [1999]:

1. The ionic current flowing is the sum of a sodium current, a potassium current, and the leak current:
$$I_{ionic} = I_{Na} + I_K + I_{leak}. \tag{3.2}$$

2. The action potential involves two major voltage-dependent ionic conductances, a sodium conductance G_{Na} and a potassium conductance G_K. The smaller "leak" conductance G_m does not depend on the membrane potential. The individual ionic currents $I_i(t)$ follow Ohm's law,
$$I_i(t) = G_i(V(t), t) \cdot (V(t) - V_{E_i}), \tag{3.3}$$
where V_{E_i} is the ionic reversal potential, given by Nernst's equation for the ionic species.

3. Each of the ionic conductance G_i is expressed by a maximum conductance, $\overline{G}_i$, multiplied by a numerical coefficient representing the fraction of the maximum conductance actually open. Hodgkin and Huxley introduced gating particles to describe the dynamics.

In the Hodgkin and Huxley model the potassium current is modeled as

$$I_K = \overline{G}_K n^4 (V - V_{E_K}) \tag{3.4}$$

where x has to be substituted by the activation n in the following equation.

$$\frac{dx}{dt} = \frac{x_\infty - x}{\tau_x} \tag{3.5}$$

with voltage-dependent time constant $\tau_n(V)$ and steady-state value $n_\infty(V)$. [Koch, 1999, Chap 6.2] motivates this equations. The sodium current is modeled by a sodium activation particle m and an inactivation particle h,

$$I_{Na} = \overline{G}_{Na} m^3 h (V - V_{E_{Na}}) \tag{3.6}$$

where m and h follow (3.5).

3.2.2 The Rothman and Manis Model for Onset Neurons

Rothman and Manis' general model for VCN neurons is an extension of the Hodgkin-Huxley model. The onset neurons are one type of those CN neurons. The onset neuron

model we discuss and characterize in this thesis is based on the work by Rothman and Manis [2003a].

In their work K^+ currents include a fast transient current I_A, a slow-inactivating low-threshold current I_{LT}, and a non-inactivating high-threshold current I_{HT}. The model also includes a sodium current I_{Na}, plus fast-inactivating high-threshold Na^+ current I_H. Analog to Equations (3.1) and (3.2) they come up with

$$C_m \frac{dV}{dt} = I_A + I_{LT} + I_{HT} + I_{Na} + I_H + I_{lk} + I_E, \tag{3.7}$$

where I_{lk} is the leakage current and I_E is the excitatory post-synaptic current. The capacity of the cell membrane is $C_m = 12$ pF.

The currents have voltage and time dependencies similar to those of the original Hodgkin and Huxley model. They are governed by activation/inactivation variables a, b, c, w, z, n, m, h and r which follow the first order differential equation as defined in (3.5). The exact equations with peak i conductance $\overline{g}_i$ are stated below; reversal potentials are: $V_k = -70$ mV, $V_{Na} = +55$ mV, $V_h = -43$ mV, and $V_{lk} = -65$ mV.

Fast transient K^+ current

$$I_A = \overline{g}_A \cdot a^4 bc \cdot (V - V_k) \tag{3.8}$$

$$a_\infty = [1 + \exp(-(V + 31)/6)]^{-1/4} \tag{3.9}$$

$$b_\infty = c_\infty = [1 + \exp((V + 66)/7)]^{-1/2} \tag{3.10}$$

$$\tau_a = 100 \cdot [7 \exp((V + 60)/14) + 29 \exp(-(V + 60)/24)]^{-1} + 0.1 \tag{3.11}$$

$$\tau_b = 1000 \cdot [14 \exp((V + 60)/27) + 29 \exp(-(V + 60)/24)]^{-1} + 1 \tag{3.12}$$

$$\tau_c = 90 \cdot [1 + \exp(-(V + 66)/17)]^{-1} + 10 \tag{3.13}$$

Low-threshold K^+ current

$$I_{LT} = \overline{g}_{LT} \cdot w^4 z \cdot (V - V_K) \tag{3.14}$$

$$w_\infty = [1 + \exp(-(V + 48)/6)]^{-1/4} \tag{3.15}$$

$$z_\infty = (1 - \zeta)\,[1 + \exp((V + 71)/10)]^{-1} + \zeta, \text{ with } \zeta = 0.5. \tag{3.16}$$

$\zeta = 0.5$ indicates that the inactivation is only partial (50%).

$$\tau_w = 100 \cdot [6 \exp((V + 60)/6) + 16 \exp(-(V + 60)/45)]^{-1} + 1.5 \tag{3.17}$$

$$\tau_z = 1000 \cdot [\exp((V + 60)/20) + \exp(-(V + 60)/8)]^{-1} + 50 \tag{3.18}$$

High-threshold K^+ current

$$I_{HT} = \overline{g}_{HT} \cdot [\varphi n^2 + (1-\varphi)p] \cdot (V - V_K), \text{ with} \varphi = 0.85 \tag{3.19}$$

$$n_\infty = [1 + \exp(-(V+15)/5)]^{-1/2} \tag{3.20}$$

$$p_\infty = [1 + \exp(-(V+23)/6)]^{-1} \tag{3.21}$$

$$\tau_n = 100 \cdot [11 \exp((V+60)/24) + 21 \exp(-(V+60)/23)]^{-1} + 0.7 \tag{3.22}$$

$$\tau_p = 100 \cdot [4 \exp((V+60)/32) + 5 \exp(-(V+60)/22)]^{-1} + 5 \tag{3.23}$$

Fast Na^+ current

$$I_{Na} = \overline{g}_{Na} \cdot m^3 h \cdot (V - V_{Na}) \tag{3.24}$$

$$m_\infty = [1 + \exp(-(V+38)/7)]^{-1} \tag{3.25}$$

$$h_\infty = [1 + \exp(-(V+65)/6)]^{-1} \tag{3.26}$$

$$\tau_m = 10 \cdot [5 \exp((V+60)/18) + 36 \exp(-(V+60)/25)]^{-1} + 0.04 \tag{3.27}$$

$$\tau_h = 100 \cdot [7 \exp((V+60)/11) + 10 \exp(-(V+60)/25)]^{-1} + 0.6 \tag{3.28}$$

High threshold Na^+ current

$$I_H = \overline{g}_H \cdot r \cdot (V - V_H) \tag{3.29}$$

$$r_\infty = [1 + \exp(-(V+76)/7)]^{-1} \tag{3.30}$$

$$\tau_r = 10^5 \cdot [237 \exp((V+60)/12) + 17 \exp(-(V+60)/14)]^{-1} + 25 \tag{3.31}$$

Leak current

$$I_{lk} = \overline{g}_{lk} \cdot r \cdot (V - V_{lk}) \tag{3.32}$$

The octopus cells includes the following ion channels: the potassium current included I_{LT} and I_{HT} , but not the I_A current. The sodium current included the I_{Na} current but not the high-threshold current I_H. Further, the model includes a mixed cation current I_h and the leak current I_{lk}.

Hyperpolarization-activated, mixed-cation current

This current is based on a measurement by Bal and Oertel [2000] and was implemented like in Manis [2004], including the temperature correction.

$$I_h = \overline{g}_h \cdot r \cdot (V - V_h) \tag{3.33}$$

$$r_\infty = 1/\left[1 + \exp\left(V + 66\right)/7\right] \tag{3.34}$$

$$\alpha = \exp\left[3 \cdot (V + 84) \cdot 96480/(8.315 \cdot (273.16 + T))\right] \tag{3.35}$$

$$\beta = \exp\left[1.8 \cdot (V + 84) \cdot 96480/(8.315 \cdot (273.16 + T))\right] \tag{3.36}$$

$$\tau_r = 0.001 \cdot \beta/\left[4.5^{(T-33)/10} \cdot 0.0029 \cdot (1 + \alpha)\right], \tag{3.37}$$

Excitatory post-synaptic current

The excitatory post synaptic currents (EPSCs) from the ANF is modeled as

$$I_E = \sum_{i-1}^{N} g_{E_i} \cdot (V_i - V_{E_E}). \tag{3.38}$$

Where V_i is the input of nerve fiber i and the reversal point $V_{E_E} = 0$. There are N ANF inputs. Each spike triggers an exponentilly decaying input-conductance g_{E_i}

$$g_{E_i} = \begin{cases} g_{E_0} \cdot w_i \exp\left(\frac{t-t_i}{\tau_E}\right) & \text{for } t > t_i. \\ 0 & \text{for else.} \end{cases} \tag{3.39}$$

where g_{E_0} determines the peak conductance, t_i is the spiking time of the i's ANF. For onset cells the decay time is $\tau_E = 0.1$ ms (38 ^{o}C).

Model Parameters

In Rothman and Manis [2003a] there are five different model configurations investigated: the Type I-c, Type I-t, Type I-II, Type II-I, and Type II-o model. The octopus neurons in this thesis are Type II-o neurons. Their parameters are given below. The unit for conductance is nS.

$\overline{g}_{Na}$	$\overline{g}_{HT}$	$\overline{g}_{LT}$	$\overline{g}_A$	$\overline{g}_h$	$\overline{g}_{lk}$	$\overline{g}_{E_0}$ @ 38^{o}C	C_M
1000	150	600	0	40	2	24.5	12 pF

3.2.3 Implementation

The model was implemented numerically in MATLAB/Simulink® and in C. We calculated the parameters a, b, c, w, z, n, m, h and r using difference equations instead of the differential equation (3.5) and iterated them in the time domain. We used conductance values and time constants corrected for a body temperature of 38°.

We stimulated the octopus neurons with 60 sub-threshold auditory nerve fibers from the auditory model as suggested by Oertel et al. [2000]. Only excitatory synapses were considered. The wide innervation of octopus neurons, covering 2–3 octaves [Willott and Bross, 1990], was not included here (In our previous work, we found that it is possible to incorporate this effect using Hebbian learning [Wang, 2003, van Hemmen, 2001]).

3.3 Results

3.3.1 Response to Injected Step Currents

The easiest way to characterize the output of the octopus neurons is to inject step currents (I_E). Figure 3.2 shows the octopus membrane potential (panel b) in response to an injected current (panel a). The octopus neuron responds with a single action potential to the first step of the current (3.5 nA, starting at $t = 0$ ms), and a steady-state depolarization, but no further action potentials. The same pattern is repeated for the second step increase in current, where the current is increased four times (14 nA, starting at $t = 10$ ms). Figure 3.2c shows the low-threshold potassium current mainly responsible for the onset response. This potassium current is unusually large in the octopus neuron [Bal and Oertel, 2001], and contributes to the cell's input conductance already at rest. In response to the input current, the potassium channels opens and effectively shunt the cell membrane, thus increasing the input current needed to drive the cell to its firing threshold. At the current offset (at $t = 20$ ms) the octopus cell undershoots the resting potential due to the deactivation of the potassium conductance.

3.3.2 Response to Pure Tone Stimuli

Figure 3.3 shows ANF and octopus neuron responses to an 1.5 kHz pure tone with stepwise increasing amplitude; sound pressure levels were 50, 70 and 90 dB (rms). The top trace shows the signal. Note that signal amplitude at 50 dB (starting at $t = 0$ s) is a factor 100 smaller than at 90 dB. The reaction of an ANF with a characteristic frequency of 1.4 kHz, slightly lower than the test tone, is displayed in the middle trace. Before a signal is applied ($t < 0$ s), the ANF fires with its spontaneous rate of approximately 30 spikes/s. When the signal is switched on, the ANF reacts with a sharp rise of its firing rate which decays again to a sustained rate of approximately 180 spikes/s. For a tenfold increase in stimulus amplitude (70 dB, $t = 50$ ms) another transient is generated and the spontaneous

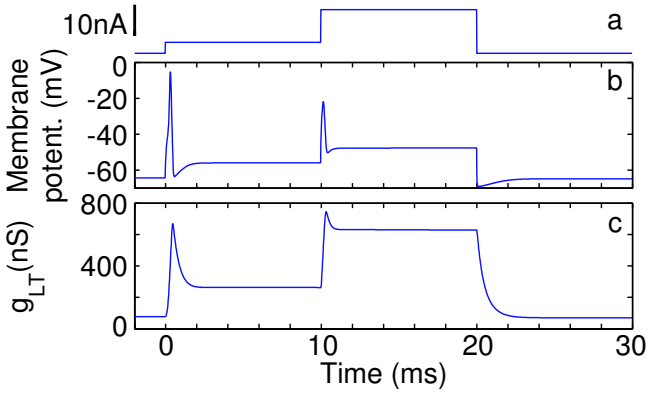

Figure 3.2: Octopus neuron response to an injected current. (a) A step current was injected into the modeled octopus neuron (0–10 ms: 3.5 nA and 10–20 ms:14 nA). (b) Octopus neuron membrane potential in response to the step current and (c) Conductance of the low-threshold potassium current (I_{LT}).

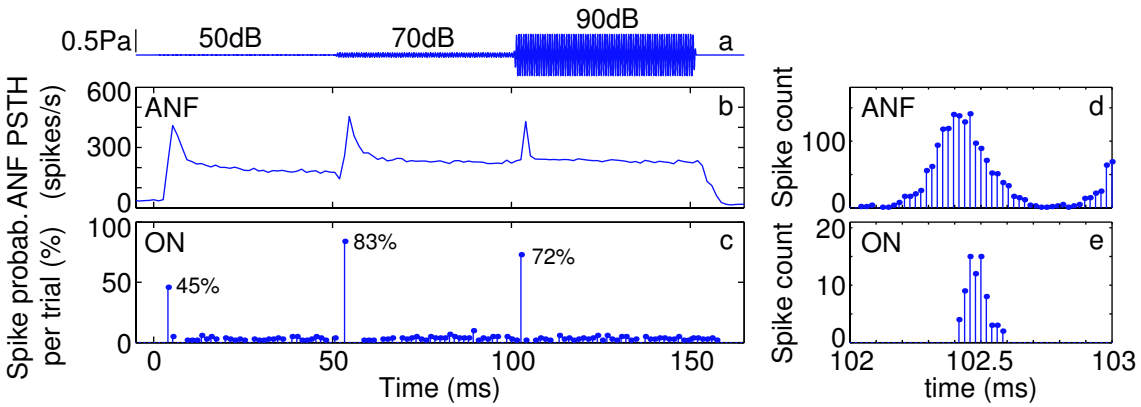

Figure 3.3: Onset processing of octopus neurons. (a) Sound stimulus is a 1.5 kHz pure tone with stepwise increasing amplitude (rise time: 1 ms). (b) Poststimulus time histogram (PSTH) of a single ANF. (c) octopus neuron activity per trial. In the left column (b+c), averaged spike counts are collected in 1.33 ms time bins, in the right column (d+e), raw spike counts are displayed for each sampling time (21 μs). Spike counts are from 60 ANFs innervating a single octopus neuron summed over 100 trials.

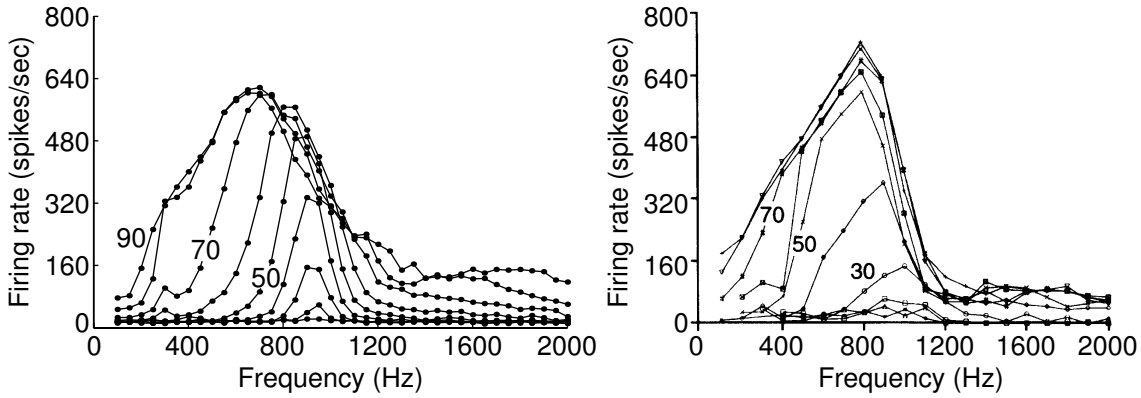

Figure 3.4: Octopus neuron responses as a function of frequency and intensity of tones. Left panel: Modeled neuron. Right panel: Measurement from a cat cochlear nucleus octopus cell [Rhode and Smith, 1986]. Both neurons had CF=950 Hz. The tones had a duration of 25 ms and levels between 10 dB and 90 dB.

rate increases to 215 spikes/s. For yet another tenfold amplitude increment the sustained firing rate hardly increases any more (220 spikes/s); HSR fibers saturate about 40 dB above threshold. Still, the fiber is able to generate a transient onset response. In Figure 3.3d the response for the step from 70 to 90 dB is shown with high temporal resolution. The stimulus changes to 90 dB at $t = 100$ ms, the change in the response of the ANF is delayed in the inner ear and occurs after approximately 102.5 ms. The responses are phase-locked, ANF activity is limited to the negative half of the stimulus cycle (rarefaction).

For high frequency sounds, octopus neurons fire only at signal onset. Figure 3.3c shows the reaction of an octopus neuron to a tone with stepwise increasing amplitude. At the onset of the 50 dB tone, the neuron firres with 31% probability. A classical integrate-and-fire neuron would respond during the whole period of the tone burst – not so for octopus neurons. During the tone burst, the excitatory input currents are shunted which prevents almost all further action potentials. This mechanism still works when the signal amplitude is increased 10-fold (even twice). When the signal level increased from 50 to 70 dB, the neuron fired with a probability of 83%, for the increase to 90 dB an action potential was elicited with 71% probability. Thereby the temporal precision of the action potentials are remarkable: whereas ANFs fire for the whole half stimulus cycle (Figure 3.3d), all of the action potentials of the octopus neuron were in an interval of less than 200 μs (Figure 3.3e). This precision is reached by coincidence detection of at least six synchronously firing ANFs and by the fast membrane time constant of octopus neurons.

Figure 3.4 compares the growth area of a modeled octopus neuron with a measurement from cat [Rhode and Smith, 1986]. Both the measured and modeled units had a CF of 950 Hz and thresholds of approximately 30 dB. Typical octopus neurons respond to tones with frequencies above approximately 800 Hz with a single well-timed spike at onset (compare Figure 3.3). For tones with frequencies below 800 Hz they can fire at every stimulus cycle, reaching average firing rates of almost 800 spikes/s! This gives the response areas of Figure 3.4 a somewhat curious look, with the strongest responses at frequencies

of 600–800 Hz, although the unit's CF might be considerably higher [Rhode and Smith, 1986]. The modeled octopus neurons do not quite reach rates of 800 Hz, but have a maximum rate of approximately 600 Hz. The shape of the response area however matches the measurement well. It should be noted that the wide response area of the octopus neuron appears although the modeled neuron was innervated by a single frequency channel only. The relatively high threshold of the octopus cells makes it respond in a level range where the auditory filters are wide. The shape of the modeled response area is more level-dependent in the mid-level range (40–70 dB) than the measured ones, more resembling the growth area of the ANF [Holmberg, 2007, Hemmert and Holmberg, 2009]. This might reflect the lack of a wide innervation pattern in the model. On the other hand, it might also reflect that the compression is stronger and filter shapes sharper in the human auditory system than in cat at low levels.

3.3.3 Response to Amplitude Modulated Signal

Among the cochlear nucleus neurons, octopus neurons are thought to best respond to amplitude modulated stimuli [Rhode, 1994]. The ability of octopus neurons to detect the envelope of the modulation signal is essential for speech processing since essentially speech signals are amplitude modulated. Octopus neurons detect the envelope of amplitude modulated signals by action potentials at a certain phase of every cycle of the modulation signal. At high frequencies where octopus neurons failed to respond to the pure tone carrier signal, they still managed to fire at each circle of the modulation signal. We discussed the distinctive ability of octopus neurons in response to AM stimuli in Chapter 2. Reverse correlation analysis (Figure 3.6) showed that they code the first derivative of the rate of ANFs, providing a good explanation of their ability to extract AM signals.

3.3.4 Response to Vowels

Vowels are characterized by periodic glottis strokes (which define their fundamental frequency). Figure 3.5 shows responses of octopus neurons to a vowel. The periodicity is clearly visible in the signal's time trace (Figure 3.5a), the output of octopus neurons (panel b),as well as in the PSTH of an auditory nerve fiber with CF 1.5 kHz (panel e). Octopus neurons extract the periodicity with great precision. Panel b can be read in a similar manner as spectrogram. The pitch frequency was revealed not only in frequency domain (by the CF of the corresponding octopus neuron) but also in time domain (the interval between neighboring spikes). At frequency regions lower than 1 kHz, the octopus neurons also show activities coding the second and third harmonics of the fundamental frequency. Several formant regions can be discerned: F1 around 500 Hz, F2-F3 around 1 kHz, and F4 around 5 kHz. The time course of the output spike trains also show the formant shift, which is typical for the dippthong "a", with F1 moving downwards in frequency with time.

The recording of the input stimulus has a ramp up at the beginning of the signal. Thus the

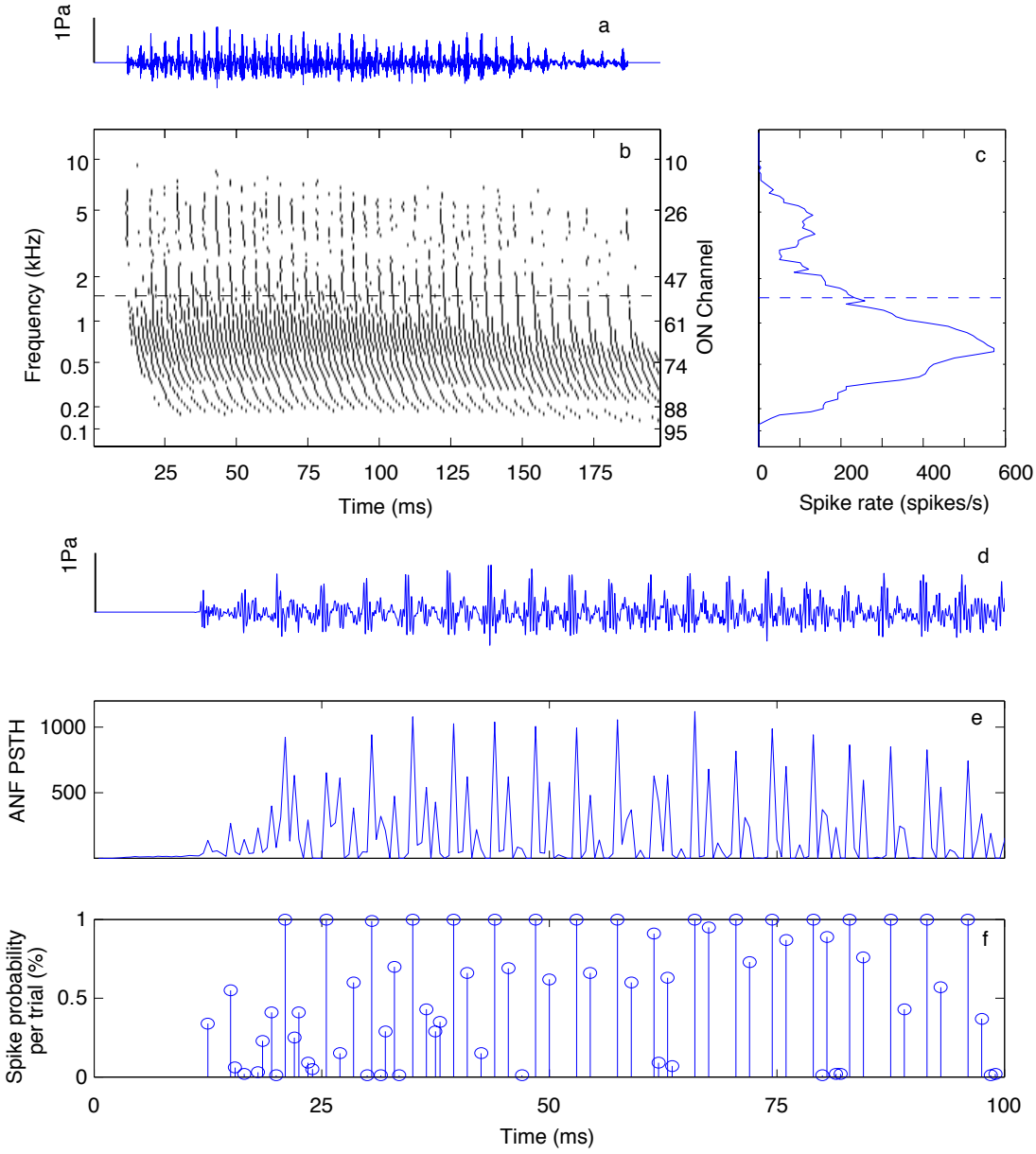

Figure 3.5: Octopus neuron response to vowel /ei/ (from ISOLET database, a female speaker with 260 Hz fundamental frequency, sound pressure level 70 dB). (a) Sound stimulus to the auditory system. (b) Outputs from 100 octopus neurons. The figure can be viewed in a similar manner as a spectrogram. (c) Average discharge rate of the octopus neurons connected to different CFs in response to the artificial vowel (no smoothing). (d) First 100 ms of the input stimulus. (e) PSTH of a single ANF (CF=1.5 kHz, 0.5 ms time bins) and (f) activity of an octopus neuron (CF=1.5 kHz, firing probabilities determined with 100 repetitions, 0.5 ms time bins).

ANFs respond at the beginning with a low rate (Figure 3.5e), which leads to a medium firing probability of the octopus neurons. After the onset, octopus neurons fire very reliably at almost each glottis stroke with high temporal precision (Firing probability was in average higher than 97%. At most of the glottis strokes, octopus neurons fire with 100% reliability. Note that we used a 0.5 ms time bin). Due to their very fast membrane time constant, octopus neurons also show intermediate activity during two neighboring glottis strokes, phase-locking to higher harmonics of the fundamental frequency (Figure 3.5f).

Panel c of Figure 3.5 shows the average discharge rate of all octopus neurons as a function of their CF. Discharge rate show several peaks which approximately correspond to different formant regions. At low frequency region around 500 Hz where entrainment takes place, octopus neurons lock to multiples of the fundamental frequency, with firing rate as high as about 600 spikes/s. At higher frequencies (1-2 kHz), they fire preferably at each glottis stroke. At very high frequency, they could not manage to fire every glottis stroke. However across channels (panel b), the one-spike-per-cycle pattern can be clearly seen.

3.3.5 Analysis with Reverse Correlation Technique

To obtain a deeper insight into how the octopus neurons work, the reverse correlation technique is applied. It reveals the temporal excitation pattern which most likely causes the octopus neurons to fire Svirskis et al. [2004].

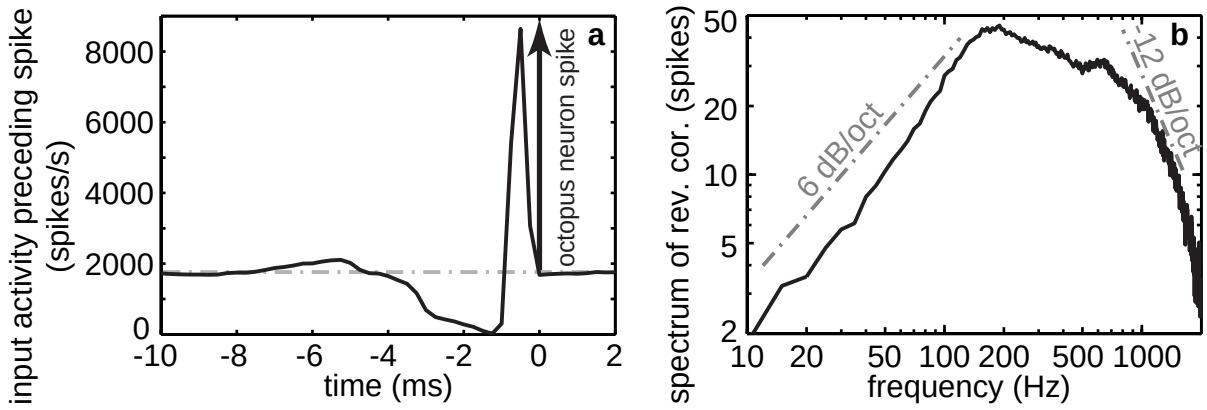

Figure 3.6: **Spike-triggered reverse-correlation and its frequency transformation.**

For the reverse correlation technique, an arbitrary input $x(t)$ is fit to the system. Here the model was stimulated by a zero signal; thus the output of the ANFs was a random process driven by its spontaneous rate, which was used as the input $x(t)$ for the onset model. The octopus neuron model responded to spontaneous ANF activity only extremely sparsely. To provide more synchronous activity, the traces of 20 ANFs were presented three times (to yield a total of 60 inputs), and, to mimic higher firing rate, the input conductance was increased by a factor of four (corresponding to a firing rate of a single ANF of 120

spikes/s) Hemmert et al. [2005]. This generated a spiking output $y(t)$, with $i = 1 \dots N$ spikes at time t_i; one spike is defined to be a dirac impulse $\delta(t)$. Thus the output is

$$y(t) = \sum_{i=1}^{N} \delta(t - t_i). \tag{3.40}$$

The reverse correlation function $R^{rev}(t)$ is defined as

$$R^{rev}(t) = \frac{1}{N} \sum_{i=1}^{N} x(t_i - t). \tag{3.41}$$

The mirrored reverse correlation function $R^{rev}(-t)$ is of special interest. It is the average activation which generates a spike at $t = 0$.

To calculate the spike-triggered reverse-correlation, ANF spikes were averaged (0.25 ms time bins) around time windows, where the octopus neurons fired (Figure 3.6a). About 10 ms before and almost immediately after the octopus neuron elicited a spike, the reverse correlation relaxes to the spontaneous rate (1800 spikes/s for 60 input fibers). To trigger a spike, highly synchronous input activity was required, which was more than a factor of 4 higher than spontaneous activity. Moreover, depressed activity about 1–3 ms in advance facilitates the generation of an action potential. The reverse correlation can be interpreted in a similar way as the impulse response of a linear filter de Boer and de Jongh [1978]. Its spectrum revealed that temporal processing of octopus neurons very much resembles a band-pass filter. The low-frequency slope was close to 6 dB/oct, the pass-band (-3 dB) reached from 110 Hz to about 650 Hz. This indicates that octopus neurons play a role in processing of amplitude modulated sounds Hemmert et al. [2005].

3.4 Discussion

In this chapter, we introduced a single compartment model of onset neurons with Hodgkin-Huxley ion channels. We used parameters from measurements to model the so called octopus neurons. However the model can be easily extended to model other types of onset neurons found in the cochlear nucleus.

Octopus neurons are interesting in speech coding, due to their suppression of spontaneous activity and steady-state stimuli and their excellent ability of coding periodic stimuli such as AM signals and voiced speech. Not less interesting is their distinctive temporal processing ability, which might play a very important role in many auditory tasks. Octopus neurons receive inputs from large number of ANFs and forward spectral and temporal information coded by ANFs to high neural processing stages with temporally sparse but precise spike trains. The information loss – or to phrase it positively, the information compression – during this transition is irreversible. It is therefore necessary to understand the properties of octopus neurons in response to different stimuli.

Experiments showed that they fire preferably at the onset of input stimuli. Even when the auditory nerve fibers stimulating them are at saturation, a step increase in stimulus

level led to reliable firing at the onset and very low or no activity at sustained input level thereafter (see Figure 3.3). In response to pure tone stimulus, octopus neurons show very strong phase-locking ability. Due to their fast membrane time constant, they are able to phase lock up to 800 Hz, which corresponds to a firing rate of 800 spikes/s. The model in this chapter achieved about 600 spikes/s in the first formant region in response to a speech stimulus. In Chapter 2, the modeled octopus neuron showed phase locking up to about 800 Hz, thanks to the realistic offset adaptation that we previously introduced. The response to pure tones with higher frequencies looked similar as those to constant signal, i.e., they fire only at the onset of the signal and were not fast enough to phase-lock to the stimuli after the onset. However, they still respond reliably at high frequency regions to amplitude modulated signals. In Chapter 2, we provided AM signals with different carrier frequencies as the input. The output spike trains of octopus neurons indicate that they phase lock to the carrier signal at low carrier frequencies (the amplitude modulation of signal is partly coded in the firing probabilities of the spikes), and mainly extract the modulation signal when carrier frequencies are high.

Given the strong coding of amplitude modulated signals, it is no surprise that the octopus neurons code the fundamental frequency of vowel stimuli very well. The direct coding of the first formant in the rate of octopus neurons is particularly interesting for speech recognition in noise and auditory scene analysis. Holmberg et al. [2007] found that the rate-place coding of vowel by ANFs at low frequencies is vulnerable to the increase of sound level but is well coded in the phase-locking of the ANFs. The temporal pattern coded by the ANFs at low frequency is transformed to a rate code in the octopus neurons, which will be transferred further. Outputs from octopus neurons at different CFs can be read as a spectrogram. It is in general possible to discern different formant regions and the shift of formants over time, which suggests that these output can be deployed for speech recognition.

We also analyzed the octopus neurons using spike triggered reverse correlation technique. Results showed that the preferred input for octopus neurons has a short silence period where the ANF rate is lower than spontaneous activity, followed by a strong and fast onset. Since there are no negative number of spikes, the spontaneous activity allows the model to have an offset by having silent period or rates that are lower than spontaneous rate. Spectrum of the reverse correlation shows a band-pass behavior with a low frequency slope of 6 dB/oct. From a signal processing point of view, the 6 dB/oct low frequency slope indicates that octopus neurons roughly work as a first order high pass filter and perform a temporal differentiation of their input (see also [Ferragamo and Oertel, 2002a]). Experiments with AM signals in Chapter 2 (Figure 2.7) also showed similar results.

The octopus neuron model was implemented mathematically in MATLAB/Simulink® and C, which are also available to other researchers.

4 Quantify Speech Information in Spike Trains using Automatic Speech Recognition

Abstract[1)]

In this chapter, we qualitatively evaluated the speech information coded by the auditory nerve fibers as well as the octopus neurons. We tested speech coding using a complex but realistic scenario: speech in noise. We compared speech recognition result using the model that we have developed before to point out the effect of modeling effort on the recognition performance: enhanced offset adaptation improves robustness of speech recognition in noisy environments. We used different acoustic modeling approaches, and concluded that multi-layer perceptrons should be considered by researchers who use auditory based feature for speech recognition. We also compared the performance of our model with human performance to show the robustness of the feature extraction by the developed auditory system at different sound levels.

[1)] Some of the speech recognition results have been published in Interspeech 2008 [Wang et al., 2008].

4.1 Introduction

Our notion of how speech is processed is still very much dominated by von Helmholtz's theory of hearing [von Helmholtz, 1863]. His notion that human ear analyzes individual frequency components of sound signals has motivated the widespread use of magnitude spectra in sound processing technologies. Strongly influenced by work from Ohm [1843], von Helmholtz put forward place coding as the principle mechanism of hearing. In a rate-place code, the spectrum is described by the profile of the discharge rates of single fibers as a function of their characteristic frequency (CF). In this chapter, we applied the rate-place code as a coding strategy to convert the output spikes of the modeled auditory system and qualitatively evaluate the information coded by auditory nerve fibers as well as octopus neurons.

Many paper have studied the encoding of speech in the auditory nerve from various aspects. Most of the early work focused on the encoding of steady-state vowels [e.g., Delgutte and Kiang, 1984a, Sachs and Young, 1979, Young and Sachs, 1979]. Later studies have included synthetic consonant-vowel syllables [e.g., Miller and Sachs, 1983, Sinex and Geisler, 1983], spoken syllables [e.g., Carney and Geisler, 1986, Deng and Geisler, 1987b], and various levels of background noise [e.g., Delgutte and Kiang, 1984b, Sachs et al., 1983, Silkes and Geisler, 1991]. Many other early works focused on the robustness of auditory encoding over a large dynamic range. Sachs and Young [1979] collected responses of hundreds of fibers to three different artificial steady-state vowels in their pioneering and widely influential experiments with speech in the cat's auditory nerve. They concluded that a rate-place representation is not robust over the range of levels encoded by humans. Due to nonlinear effects such as discharge rate saturation and suppression, there are no formant related peaks in the rate profiles at normal conversation levels (vowel levels above 65 dB(SPL). A number of more recent studies have suggested that rate-place coding of vowels might in fact be sufficient to explain the dynamic range of human vowel perception. Conley and Keilson [1995] studied discrimination of second formant frequencies in the cat's auditory nerve. When looking at the response of the whole auditory nerve and in particular the rate difference between two vowels, the formant peaks were clearly represented, even at sound levels of 70 dB(SPL). These previous studies lead to very useful results and good knowledge about the property of auditory coding, such as discrimination ability, robustness. The draw back of most of the *in vivo* studies is that there is no qualitative measure of how well a hypothetical speech encoding really works. Also, discrimination tasks are very limited as they studied only well-defined properties (e.g. formants of the speech).

In this chapter, we exploited automatic speech recognition as a tool to qualitatively evaluate the principle of rate-place code. We tested speech coding in the auditory system with a complex but realistic approach: speech in noise. We used the previously developed model of human auditory processing up to the level of the auditory nerve to further investigate the ability of rate-place code and to explain human auditory perception. Compared to previous discrimination tasks, automatic speech recognition outperforms clearly in terms of generality. Speech recognition enables complex tasks with various stimuli and can eas-

ily cope with many variations such as different speakers, different sound levels and various noise scenarios. After all, speech coding aims to serve the perception of speech. Therefore automatic speech recognition is genuinely a very good and straight-forward evaluation approach to qualitatively assess the performance of any speech coding strategy.

Several researchers have tried auditory models as front end in ASR. The focus has however rather been to try to improve robustness of ASR by using principles found in the auditory system, than testing the coding strategies per se. Following the findings of Young and Sachs [1979], most have used various types of temporal coding of speech. Auditory models previously used in ASR include those of Seneff [1986, 1988], Ghitza [1994], and Sheikhzadeh and Deng [1998]. Most models have not deployed actual spike trains. The ensemble interval histogram (EIH) model [Ghitza, 1994] did generate spikes, although it used a very simple deterministic zero-crossing algorithm. Ghitza's underlying BM model was adopted from Goldstein [1990], and modeled compression as well as suppression effects. The model of Lazzaro and Wawrzynek [1997] also generated spikes. They used a silicon cochlea model in an ASR task. Their model consisted of a linear BM model, and an inner hair cell (IHC) and auditory nerve (AN) model. The latter part of the model performed an amplitude compression, half-wave rectification and a conversion into probabilistic trains of spikes. However, the auditory filters did not change with level. Compared to earlier auditory models used in ASR, the model used in this thesis is unique in some important respects: it reproduces the latest estimates of auditory filters in human auditory processing – which are considerably more sharply tuned than animal data suggests [Shera et al., 2002]. It shows frequency tuning that closely resembles human psychoacoustic estimations over a wide range of levels. It generates spikes in a fashion closely resembling the real auditory system. It also includes both low- and high spontaneous rate fibers.

In our work, we used the framework of automatic speech recognition with Hidden Markov Models (HMMs) as the back end. We extracted features for speech recognition by post processing the output spikes from the model of human auditory processing. Later we augmented this framework using the hybrid connectionist approach, where multi-layer perceptrons (MLPs) are used for acoustic modeling.

In the following part of this chapter, we will first give a brief introduction to HMM based speech recognition, the hybrid connectionist approach and the speech recognition task. Then we will show how we use the model of human auditory processing to generate features and how we input these features into the speech recognition engine. The test consists of the English alphabets, with and without background noise. Features from auditory nerve fibers and octopus neurons separately as well as the combination of both were tested. It is shown that offset adaptation not only makes the output spikes of auditory fibers and octopus neuron more realistic, but also led to improved speech recognition performance per se. We analyzed the dependance of speech recognition on input sound levels and compared modeled results with human performances. Recognition performance using different auditory modeling methods were also compared.

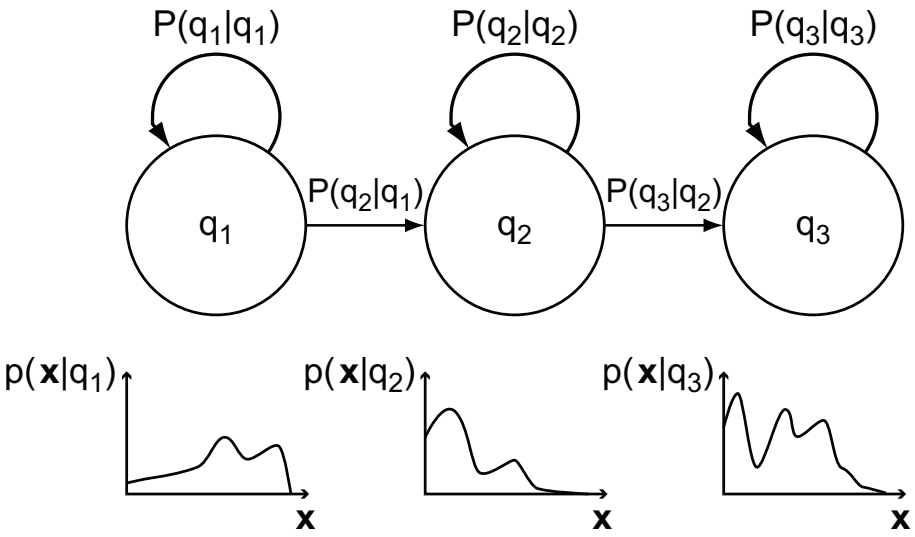

Figure 4.1: A schematic representation of a Hidden Markov Model (HMM). An HMM is defined by the states q_i, the allowed transition probabilities $P(q_j|q_i)$, and the emission probability densities $p(\mathbf{x}|q_i)$, where $\mathbf{x}$ refers to the multidimensional feature space.

4.2 Speech Recognition with HMMs

Figure 4.1 shows a schematic representation of a Hidden Markov Model. HMMs have a number of advantages when investigating the amount of speech information contained in the rate-place code. HMMs are data-driven, and thus few assumptions have to be made on the nature of classification. In particular, HMMs handle time-warping in a data-driven manner. This enables more complex tests. Stimuli can have varying durations and include varying numbers of phonemes. Variability such as different speakers and background noise can easily be incorporated into the classification tasks. An HMM approach to measure the information in the rate-place code also inherits the limitations of HMMs, such as the assumption of independently and identically distributed acoustic vectors within the speech segment corresponding to a state and the assumed correctness of the acoustic models [Morgan and Bourlard, 1995]. While HMMs are the standard approach to ASR today, many in the ASR field believe that the current gap between human and ASR performance is in large part due to the limitations of the HMMs. It has been argued that the statistical assumptions typically incorporated into HMMs used in speech recognition may constrain their ability to exploit available information [Holmes and Russell, 1999]. Also, the degree of success with which an HMM can model speech depends on the amount of training data available and how well training conditions match later observations.

4.3 Speech Recognition with Multi-Layer Perceptrons

Statistical representations using Hidden Markov Models have been developed to the point where impressive recognition performance can be achieved in the laboratory on very large vocabulary speaker-independent continuous speech recognition tasks. The HMM Tool Kit from Cambridge has been widely spread so that researchers can build up speech recognition systems based on some of the best known statistical approaches. In general, HMM-based structures and algorithms provide a rich and flexible mathematical framework for building recognition systems. They also feature powerful learning and decoding methods for temporal sequences without requiring any explicit segmentation in terms of the speech units (typically phones or phonemes) used, which serves as a basis for continuous speech training and recognition. Also, they easily accommodate different levels of constraints, such as phonological and syntactical constraints, as long as they are expressed in terms of the same statistical formalism. HMM based structures and algorithms have very strong representation power. However, to take advantage of this representation power, algorithms must explicitly or implicitly make numerous assumptions about speech, some of which obviously unrealistic. For instance, it is assumed that there is no correlation between input features and the input features can be parameterized as mixtures of Gaussian densities. It may be the case that some of the characteristics of HMM based approaches to speech recognition are limiting factors for further improvement in the long run.

Therefore in this section, we evaluate speech recognition performance with a hybrid connectionist approach. In the connectionist approach, multi-layer perceptrons (MLPs) are used for acoustic modeling. MLP acoustic modeling provides one way to reduce system dependency on unrealistic assumptions about speech, and sometimes gives better performance than Gaussian mixture model (GMM) acoustic modeling when working with novel feature vectors. Another advantage of MLP acoustic modeling is that it allows for a straightforward and effective multi-stream speech recognition approach. Therefore we can combine different features with great ease. In the multi-stream approach, a separate MLP is used for each stream; streams are then combined by combining posterior probabilities at the frame level.

Typically, MLPs have a layered feed forward architecture with an input layer (consisting of the input variables), zero or more hidden (intermediate) layers, and an output layer. Each layer computes a set of linear discrimination functions via a weight matrix followed by a nonlinear function, which is often a sigmoid function. In principle, MLPs with enough hidden units can provide arbitrary mappings between the inputs and the outputs. MLP parameters – the element of the weight matrices – are trained to associate a "desired" output vector with an input vector. However, it is not the best practice to use MLP only for speech recognition tasks, especially continuous speech recognition. MLPs, or more generally, Artificial Neural Networks (ANNs), can be used to classify speech units such as phonemes or words, typically by mapping temporal representations into spacial ones. However, ANNs classifying complete temporal sequences haven't been successful for continuous speech recognition. In fact, they are doomed not to work well for continuous

speech since the number of possible word sequences in an utterance is generally infinite. There hasn't been any known principled way to translate an input sequence of acoustic vectors into an output sequence of speech units with an ANN only. On the other hand, HMMs provide a reasonable structure for representing sequences of speech sounds or words, as we previously mentioned. Given such a structure, one good use for ANNs is to provide the local distance measure. In particular, given the HMM system, we would like to estimate the emission probability $p(x_n|q_k)$ as in Figure 4.1, that is, the probability of the observed data vector given the hypothesized HMM state, which corresponds to some speech sound. However, HMMs are based on a very strict formalism that is difficult to modify without losing the theoretical foundations or the efficiency of the training and recognition algorithms. Fortunately, ANNs can estimate probabilities that are related to the emission probabilities, thus can be easily integrated into an HMM-based approach.

It has been exprimentally observed that, for systems trained on a large amount of speech, the outputs of properly trained MLP approximate posterior probabilities ($p(q_k|x_n)$. Thus emission probabilities ($x_n|q_k$) can be estimated by applying Bayer's rule to the MLP outputs [Morgan and Bourlard, 1995].

$$\frac{p(x_n|q_k)}{p(x_n)} = \frac{p(q_k|x_n)}{p(q_k)} \tag{4.1}$$

Figure 4.2 shows the basic hybrid scheme, in which the MLP generates posterior estimates that can be transformed into emission probabilities as described above, and then used in dynamic programming either for forced alignment (when the word sequence is assumed) or for recognition (when word sequences are hypothesized).

Ultimately, the connectionist recognition schema uses the same structure as HMM-based schema, except that we derive the same probability with MLP rather than with a conventional estimator, such as Gaussian mixture as used in HMM-based speech recognizer.

Our testbed handles MLPs using ICSI's qnstrn (training) and qnsfwd (forward pass) tools, which can be found in the Quicknet3 package. Decoding and coding (hypothesis search) are handled using the Noway decoder, available as part of the SPRACHcore package. [2)].

4.4 Automatic Speech Recognition with the Auditory Model

In the following sections, we give a brief description of the auditory model, the ASR testbed setup and the speech recognition task. We used a test set that consists of the isolate English alphabet, with and without background noise. Using the ASR setup, we tested the robustness of the auditory coding against noise. We also tested how well

[2)] For more information about these software packages, see http://www.icsi.berkeley.edu/Speech/icsi-speech-tools.html

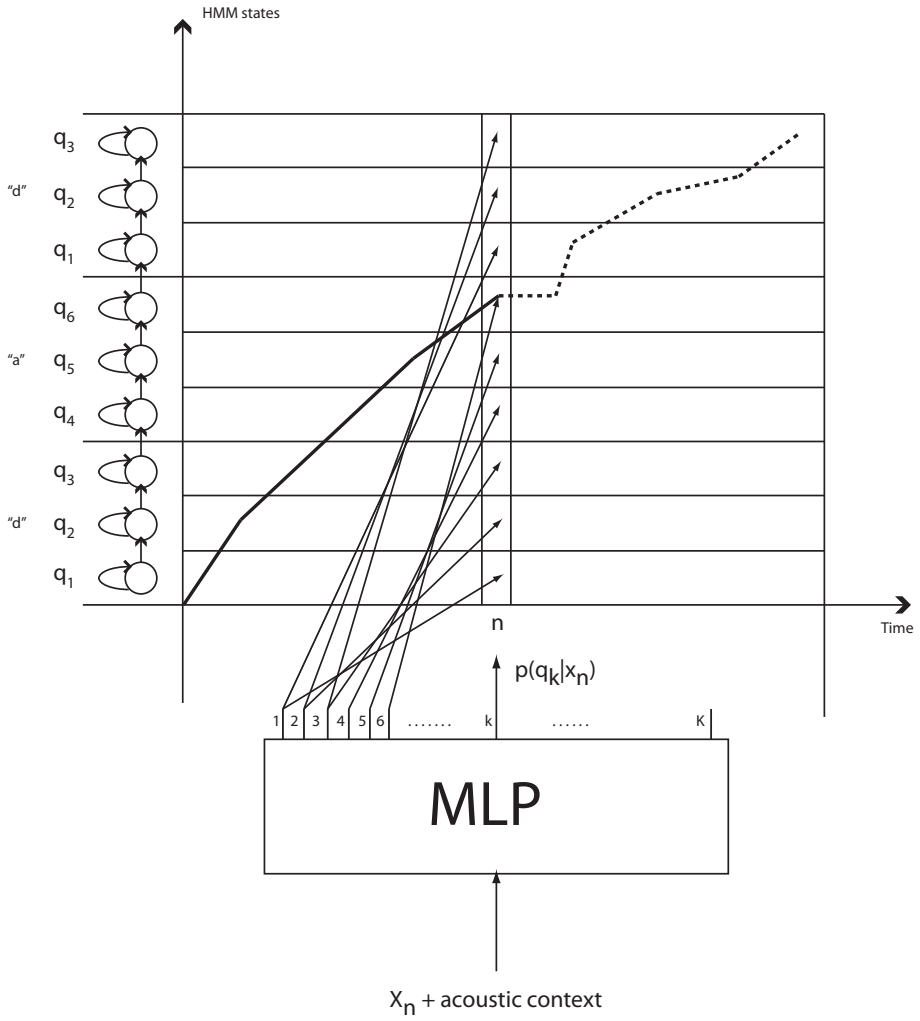

Figure 4.2: At every step n, the acoustic vector x_n with right and left conext is presented to the network of multilayer perceptrons. This generates local probabilities that are used, after division by priors, as local scaled likelihoods in a Viterbi dynamic programming algorithm. The arrows coming up from each MLP output symbolize the use of these scaled likelihoods (after taking the negative logarithm) as distances from the acoustic input to their corresponding state. The dark solid line shows the best path through the models up to time n (that can be determined by backtracking through pointers), and the dashed path shows its continuation that can be determined once the distances are computed for the last frame in the data.

a rate-place representation of the speech works at different sound levels, both in clean and noisy conditions. By using different auditory modeling (with or without enhanced offset adaptation) and different acoustic modeling, we showed the importance of realistic auditory and acoustic modeling to the recognition of speech.

4.4.1 Model and Interface for Feature Extraction

The auditory model and its interface to the HMM speech recognizer are shown in Figure 4.3. We use the model introduced in Section 1.4 with a few modifications to reduce the computational complexity. In the following sections we briefly describe the main components of the auditory model, and the interface between the model and the HMM speech recognizer.

The model of the peripheral hearing system consists of a simplified ear canal and middle ear (described in Sec. 1.4.1), a model of inner ear hydrodynamics followed by a compression stage, and an inner hair cell model. BM vibrations were calculated with a computationally efficient wave-digital filter model comprised of 100 sections at a sampling rate of 48 kHz. This inner ear model covers the complete human hearing range up to 20 kHz; as the sound samples used in this chapter are band limited to 8 kHz, we discarded 9 high-frequency channels and processed only the remaining 91 channels. The model of the sensory cells and synaptic mechanisms is based on a model by Sumner et al. [2002]. The modified model works with continuous vesicle pools instead of quantized ones to speed up calculation. Each sensory cell was connected to multiple auditory nerve fibers (ANFs), which elicited all-or-none nerve action potentials. We employed a simplified spike generation module, based on a binomial distribution of spike probabilities over the ensemble of nerve fibers. The module still included both absolute and relative refraction effects. High-spontaneous rate (HSR) and low-spontaneous rate (LSR) ANFs with different thresholds and growth function slopes were modeled (Figure 4.6). No variations in spontaneous rate or thresholds within each fiber type were modeled. The inner ear model provides large dynamic compression of more than 80 dB and generates realistic ANF responses. It replicates the bandwidths of human threshold tuning curves [Shera et al., 2002] and latest measurements of dynamic range compression [Lopez-Poveda et al., 2003] with great precision [Holmberg and Hemmert, 2004].

The right hand side of Figure 4.3 shows how the spike trains of the ANFs are interfaced to the HMM speech recognizer. The auditory nerve features are based on a rate-place coding strategy. There are a number of conceptual differences between standard ASR features and the auditory features: Whereas the spike trains of the auditory nerve give a sparse but highly precise temporal code of speech, ASR systems rely on time-averaged smooth features; the auditory system has a high frequency resolution with highly correlated channels whereas ASR features tend to be less correlated and have fewer components. Temporal averaging of the highly precise temporal code of the ANFs is necessary to generate the rate-place code and adapt the features to the ASR system. Therefore, we use a temporal resolution typical for ASR, averaging the response of each channel using a 25 ms

Hanning window advanced in steps of 10 ms. The resulting features still have a relatively high spectral resolution of 91 channels where as typical Mel-frequency cepstral coefficient [MFCC] features have a dimensionality of 12–14. Further, the different frequency channels are highly correlated, which is undesirable since we use diagonal-covariance Gaussians in the HMM (see Section 4.4.2). Therefore, we used a discrete cosine transform (DCT) to reduce the spectral resolution and decorrelate the feature vectors[3)]. We kept the first twelve cepstral coefficients, including C0. To better match the Gaussian assumptions of the HMM we gaussianized the amplitude distribution of the features, using a nonlinear transformation based on the data in the training set [4)]. In the last step, we augmented the auditory nerve fiber features by first- and second order delta coefficients using HCopy tool from HTK, calculated over four frames of past and four frames of future data. Delta coefficients are a standard component of ASR systems. They greatly improve recognition scores by introducing context information to the short-term features. In addition to deltas and double deltas, the MLP testbed also use a five-frame context window, due to its capability of processing large amount of input data. The various stages of the feature calculation are shown on the right hand side in Figure 4.3.

4.4.2 Recognition Task and Recognizer Back End

The automatic speech recognition tests presented in this chapter were carried out on the ISOLET database, and on a modified version of ISOLET with artificially added noise (noisy ISOLET). ISOLET [Cole et al., 1990] contains 150 speakers (75 female and 75 male speakers), each speaking the whole alphabet twice. The recordings were made of isolated words in a quiet environment. For the noisy ISOLET database, one of eight different noise types was artificially added to each utterance at one of six different A-weighted signal-to-noise ratios: clean, 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB. The recognition scripts and tools to reproduce the noisy ISOLET corpus that we used are available at `www.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet`. The noise types were chosen from the RSG-10 collection [Steeneken and Geurtsen, 1988]. The original division of the ISOLET database into five subsets was kept, and the setup uses five-way cross-validation to increase the statistical significance of the results [Holmberg et al., 2007]. The experiments were carried out for two different training conditions: clean training where training data was taken from the original ISOLET only, and multi-condition training with four different noise types at various SNR levels. The noise types were added in such a manner that, in case of multi-condition training and noisy test, each of the five test sets had three matched noises and one mismatched noise, i.e., a noise type not seen during training. Henceforth, we will refer to tests performed on the noisy ISOLET as noisy tests, and those performed on the original ISOLET data as clean tests. We used the first of the five splits of ISOLET database to tune SPRACHcore decoder options in

[3)] Statistics-driven approaches like Karhunen-Loève transformation (KLT) and linear discriminant analysis (LDA) were also tried. None of these alternative approaches performed consistently better than DCT on the present ASR task.

[4)] We used Jeff Bilmes' pfile_gaussian tool [Ellis, a].

MLP based speech recognition. Tuning HTK decoder options had no significant effect for this task. We reported results on the remaining four splits, 6240 words in total.

The ISOLET and noisy ISOLET recordings were scaled to have physically meaningful amplitudes. When scaling, each utterance in the corpus was scaled by the same value. Thus, the procedure scales the dynamic range of the whole set of recordings, but does not normalize the differences between single utterances. The absolute sound levels reported throughout the paper are A-weighted rms levels, averaged over all clean utterances including pauses.

We used two speech recognizers: one built with Cambridge's HTK using Gaussian mixture models (GMMs), and one built with SPRACHcore [Ellis, a] using multi-layer perceptrons (MLPs) [Morgan and Bourlard, 1995] . The recognition scripts and noisy ISOLET corpus that we used are available at www.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet.

With HTK we used six states per word (one state for the pause model) and eight diagonal-covariance Gaussians per state, and with SPRACHcore we used 1600 MLP hidden units. In earlier work (not using all the same features), we found that increasing acoustic model sizes beyond this had only a minor effect. When using HTK for the auditory features (not for MFCC features [5)]), to make the features easier for a GMM to model, we gaussianized the feature vectors prior to delta calculation, using SPRACHcore pfile_gaussian tool. Further regarding acoustic modeling, the HTK based recognizer used whole word models, one model for each letter; the MLP based recognizer instead used phoneme based models. We tried MLP acoustic modeling because we hoped avoiding the statistical assumptions of the GMM approach would provide useful flexibility when working with the model-based features [Ellis, b]. While we did try to de-correlate and gaussianize the features for the GMMs, we knew this might not be enough – for example, those transformations worked with global distributions while GMMs model state conditioned distributions.

4.4.3 Speech Recognition Baseline

We adopted the commonly used Mel Frequency Cepstrum Coefficient (MFCC) ASR features as the baseline result for comparing recognition scores. We calculated MFCC features using HCopy tool from HTK. We took twelve MFCC coefficients (C0-C11) augmented with first- and second-order derivatives calculated using four frames of past information and four frames of future information, resulting in a 36-dimensional ASR feature vector.

[5)] With gaussianization ASR results slightly improved for auditory features, but worsened a little for MFCC features. See also results for MSG features [Ellis, b].

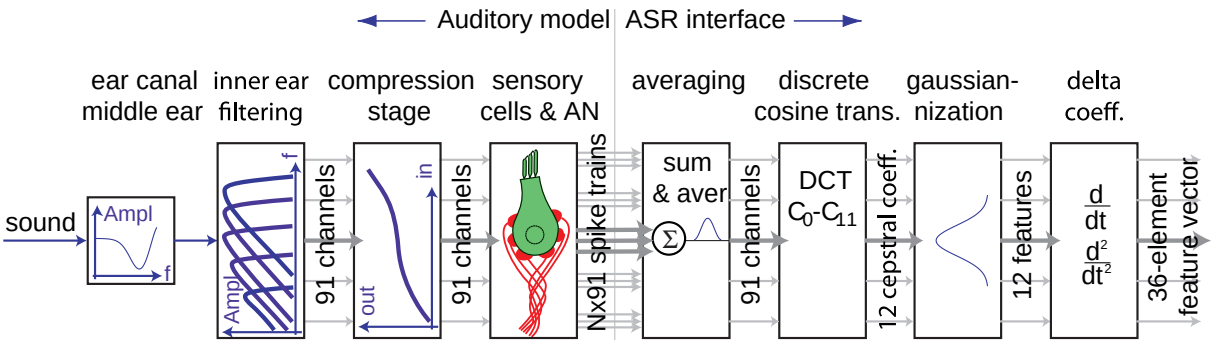

Figure 4.3: Schematic figure of the auditory model and the interface to the ASR system. Sound signals are separated into 91 frequency channels in the model of the inner ear hydrodynamics. The compression stage introduces the amplification and nonlinear characteristics of the inner ear. N auditory nerve fibers innervate one sensory cell (inner hair cell). The ANFs can be high or low spontaneous rate fibers. To interface to the ASR system, the spike-trains were averaged using a 25 ms Hanning window, advanced in steps of 10 ms. The first twelve components of the discrete cosine transform were kept, and the resulting features' amplitude distributions were gaussianized. Finally, first- and second-order delta coefficients were added.

4.5 Augmenting the Rate Code by Cochlea Nucleus Octopus Neurons

The major difference between the human auditory system and automatic speech recognition lies in their representation of sound signals. ASR uses a smoothed low-dimensional temporal and spectral representation of sound signals whereas the hearing system relies on extremely high-dimensional but temporally sparse spike trains. The strength of the latter representation lies in the inherent coding of temporal information, which is exploited by neuronal networks along the auditory pathway. So far, features derived from the auditory nerve fibers represent mainly the spectral characteristics of speech signal, very much resembling the ASR systems. The precise temporal information was largely discarded by the temporal averaging. Therefore in this section, we complemented the features derived from ANFs with features derived from spike trains of octopus neurons, which are known for their distinctive temporal processing properties. Since octopus neurons primarily respond to voiced speech, we used vowels from the English alphabets (a, e, i, o, u and y) to perform speech recognition tests.

The octopus neurons are found in the posterior ventral cochlear nucleus (PVCN), and belong to the functional group of onset units [Rhode and Greenberg, 1992]. They suppress steady-state activity and fire on signal onsets. They are sensitive to amplitude modulated signals especially in the frequency range of human speech signal, and further enhance the amplitude modulation of voiced speech.

In this section, we used the average firing rates of octopus neurons to augment the rate place code of the auditory nerve fibers. The rate-place contour of octopus neurons in response to vowels is fundamentally different from that of ANFs. Octopus neurons only respond to stimuli with a clear temporal structure. Its firing rate depends only partially on the intensity of the stimuli, but is more related to the fundamental frequency and the amplitude modulation frequency (see Chapter 2.3). In contrast, ANFs respond to any input within their response area and its firing rate mainly depends on the intensity of the stimuli. Therefore the octopus neurons are capable of translating the temporal information into a rate code. Figure 4.4 shows the schematics of combing the features from ANFs and octopus neurons for speech recognition tests. We used 172,000 ANFs and 182 octopus neurons each driven by 60 ANFs. We used MLP testbed for combining the features of ANFs and octopus neurons. Feature extraction was the same as we described in previous chapter. In this multi-stream architecture, each stream of features was passed to a different neural network for phone posterior probability estimation, and the posterior probabilities are then combined across the different neural networks at the frame level. After combination, there is only a single stream of probabilities, which are passed to the decoder for recognition. For more information on this posterior combination approach to multi-stream speech recognition, please refer to Ellis [2000].

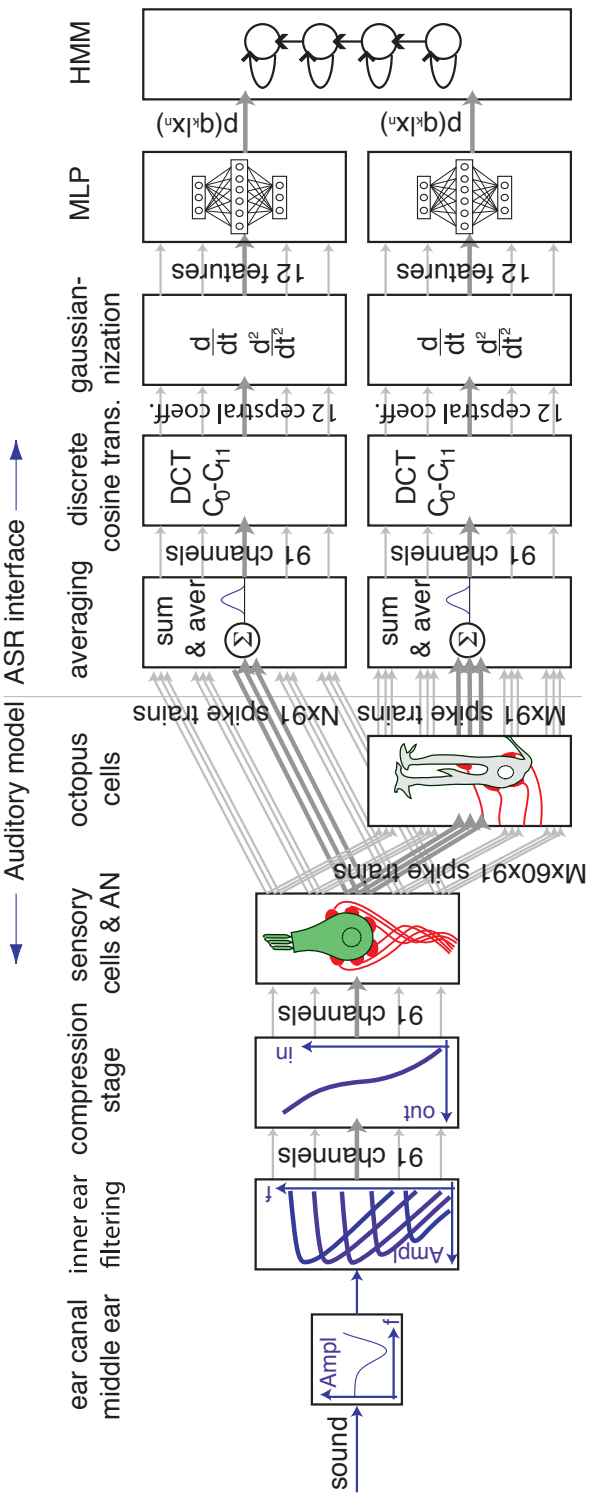

Figure 4.4: Schematic figure of the auditory model and the interface to the ASR system. Sound signals are separated into 91 frequency channels in the model of the inner ear hydrodynamics (9 high-frequency channels were discarded). The compression stage introduces the amplification and nonlinear characteristics of the inner ear. The spike trains of the auditory nerve fibers are used to calculate rate-place features, just like in Chapter 4. Here we used $N(N = 189)$ high spontaneous auditory nerve fibers per inner hair cell. The octopus neurons are innervated by a subset of the ANFs, in this case 60 HSR ANFs. There are $M(M = 2)$ octopus neurons at each CF, each innervated by a different subset of ANFs. To interface to the ASR system, the spike-trains were averaged using 25 ms Hanning windows, advanced in steps of 10 ms. For each feature type, the first twelve components of the discrete cosine transform were kept. Then the feature vectors were complemented by first- and second-order delta coefficients calculated using 4 frames in the past and 4 frames in the future. Up to this point the ANF-based features and the octopus ones were treated separately. The feature vectors were passed through different MLP networks to estimate the phone posterior probability. The posterior probabilities were then combined at the frame level across the different neural networks. After combination there is a single stream of probabilities which are then passed to the decoder for recognition.

4.6 Results

Figure 4.5 shows speech recognition results as a function of SNR using features extracted from auditory nerve (panel a) and octopus neuron (panel b) spike trains. As octopus neurons respond primarily to voiced speech, we tested them only on the vowel subset (a, e, i, o, u and y) of ISOLET.

Using MLP testbed instead of HTK testbed resulted in major performance improvements for all the auditory model based features (see also Table 4.1). The relative improvement in word error rate (WER) averaged over all SNR levels is very significant: 27.2% for ANF without OA, 33.1% for ANF with OA, 36.3% for octopus neuron without OA and 37.6% for octopus neuron with OA. All these improvements were statistically significant using a difference of portions significance test ($P < 0.0001$). With MFCC features, for the full set there was no statistically significant difference between MLP and HTK, while for the vowel subset using HTK reduced word error rate considerably by 20% (this was statistically significant, $P < 0.002$).

Including the enhanced offset adaptation model resulted in very large performance improvements for features derived from ANFs. The relative improvement in WER averaged over all SNR levels is: 12.5% with HTK testbed and 19.6% with MLP testbed.

The enhanced offset adaptation improves phase locking of ANFs, which is vital for further neuronal processing stages. Octopus neurons only responded in the frequency region above 3 kHz when offset adaptation was included in the IHC-AN model (refer to Chapter 2). The improved firing pattern of octopus neurons leads to considerable performance improvements in speech recognition on the vowel subset. The relative improvement in WER averaged over all SNR level is as high as 37.1% with HTK testbed and 38.4% with MLP testbed.

The improved speech recognition performance for both ANFs and octopus neurons showed that the enhanced offset adaptation not only provided more useful input for octopus neurons but also improved speech coding per se. All the improvements resulted from enhanced offset adaptation were statistically significant ($P < 0.0001$).

4.6.1 Level Dependency of the Speech Recognition Results

The main problem of the rate-place code is the limited dynamic range of auditory nerve fibers. Sachs and Young [1979], for example, concluded that a rate-place code representation of speech is not robust over the range of levels encoded by humans. Due to nonlinear effects such as discharge rate saturation and suppression, there are no formant related peaks in the rate profiles at normal conversational levels (vowel levels above 65 dB SPL). Figure 4.6 shows the modeled rate-level functions of two auditory nerve fibers with characteristic frequency of 1.5kHz in response to pure tone simuli. These curves match very well with physiological measurements (data not shown). HSR fibers have low threshold and a limited dynamic range of about less than 40 dB. We calculated speech recognition

		HTK	MLP
Full noisy ISOLET (17200 HSR ANFs)	MFCC	82.8	83.3
	no OA	71.3	79.1
	enhanced OA	74.9	83.2
Vowel subset (182 octopus neurons)	MFCC	95.2	94.0
	no OA	76.3	84.9
	enhanced OA	85.1	90.7

Table 4.1: Recognition results averaged over all SNR levels for multi-conditional training (OA: offset adaptation). MLP testbed outperforms HTK testbed in most of the experiments except in the case of vowel recognition using MFCC features. For the full alphabet recognition task, auditory feature with enhanced offset adaptation achieved comparable performance (WER 83.2%) as to the result of MFCC features (83.3%).

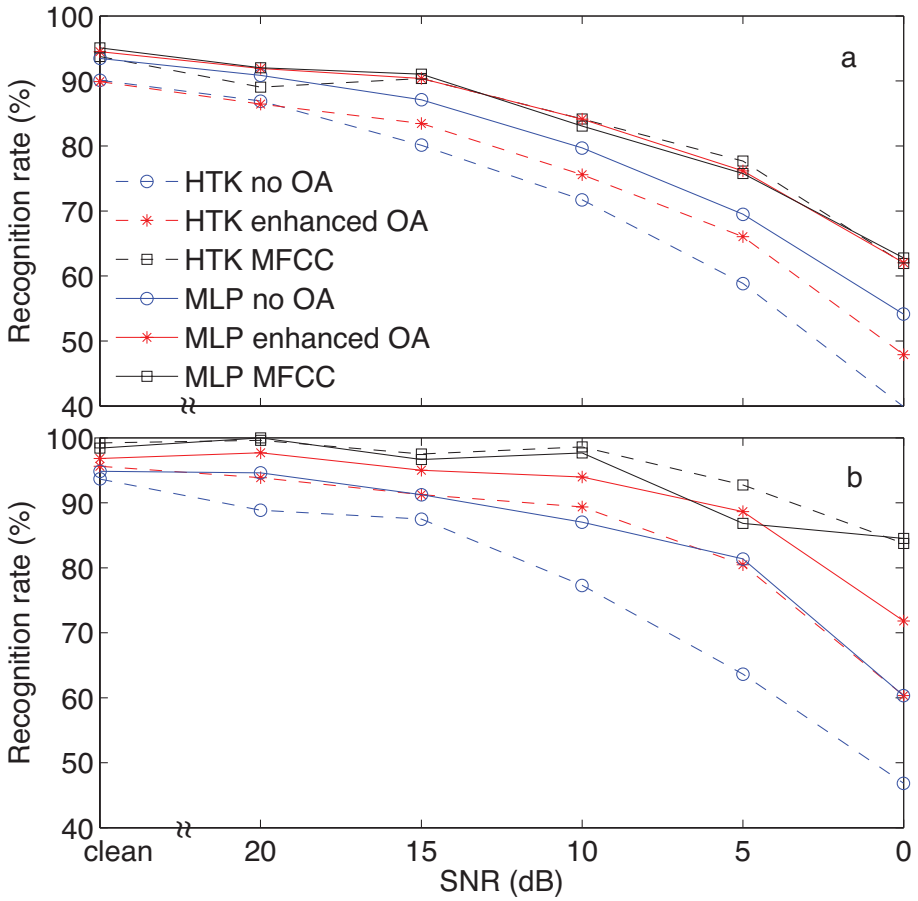

Figure 4.5: Speech recognition results as a function of SNR. Panel a shows results on the full noisy ISOLET task for features derived from 17,200 ANFs. Panel b shows results on the vowel subset (a, e, i, o, u and y) for features derived from 182 octopus neurons.

results using features derived from HSR fibers at different speech intensity to show how well the rate-place code works at different levels, in particular at high sound levels.

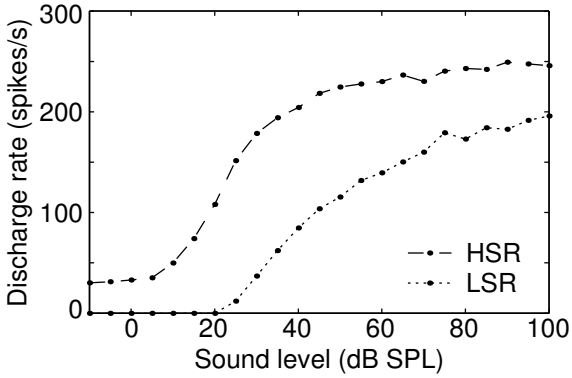

Figure 4.6: Modeled auditory nerve rate-level functions for a high spontaneous rate fiber (HSR, dashed line) and a low spontaneous rate fiber (LSR, dotted line). Both fibers had a CF of 1.5 kHz, and stimuli were pure tones at CF with 100 ms duration (100 repetitions).

Figure 4.7 shows recognition results at different sound levels for the multi-conditional training. Baseline is speech recognition results using the model without enhanced offset adaptation and HTK testbed, where reasonably good result was achieved at low sound level. At 25 dB, recognition rate was 72.7%. As sound level increased, recognition performance increased also slightly until it reached the plateau at 45 dB. Increasing sound level beyond 45 dB degraded recognition performance dramatically: recognition rate dropped at a rate of about 9.4% per 10 dB and scored only 32.3% at 105 dB SPL. After we improved the model with enhanced offset adaptation, speech recognition performance degraded slightly at low SPLs from 25 dB to about 60 dB, but improved considerably at high SPLs compared to the original model. At 105 dB SPL, the enhanced model improved recognition score by 95.7% relatively. In general, the enhanced model is much more robust to the change of SPL. The recognition score only dropped by 10.7% when SPL increased from 45 dB to 105 dB.

Using MLP based speech recognizer significantly improved performance for both models with and without enhanced offset adaptation. For the model with enhanced offset adaptation, MLP acoustic modeling increased recognition score homogeneously at each SPL of the input signals, lifting the curve for recognition score in Figure 4.7 almost parallel compared to HTK. The recognition rate averaged over all SPLs increased from 69.0% using HTK to 75.8% using MLP. The performance improvement for the model without enhanced offset adaptation was even more impressive. By using MLP, we achieved a recognition performance of 74.2% averaged over all SPLs, comparing to 61.9% using HTK. MLP alone significantly increased the robustness of the recognition performance against increasing SPL. From 25 dB to 85 dB we were able to achieve stable recognition rates. However, SPLs stronger than 85 dB deteriorated the recognition performance, with a rate of about 9.6% per 10 dB, similar to what we calculated previously when using HTK.

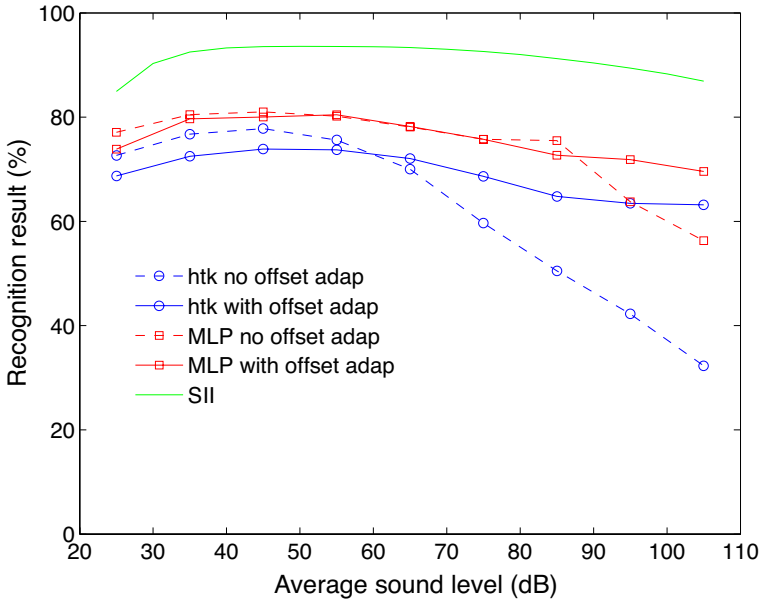

Figure 4.7: Recognition results at different speech level from 15 dB SPL(A) to 105 dB SPL(A) in 10 dB step. Results for models with and without enhanced OA were calculated using both MLP and HTK based recognizer. Multi-conditional training was used for the experiments. SII refers to speech intelligence index which represents the human performance (see Section 4.6.3).

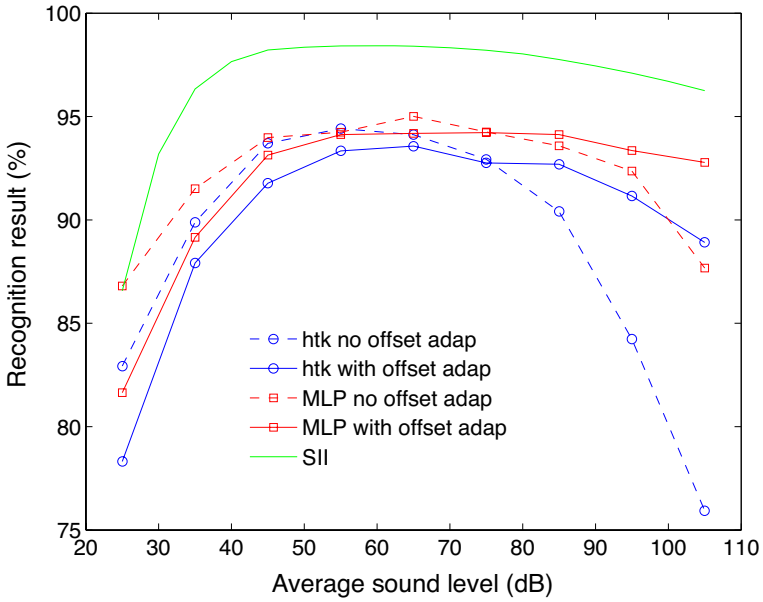

Figure 4.8: Recognition results at different speech level from 15 dB SPL(A) to 105 dB SPL(A) in 10 dB step. Results for models with and without enhanced OA were calculated using both MLP and HTK based recognizer. Clean condition was used for the experiments.

It is worth noticing that by using MLP instead of HTK only, the speech recognition performance was improved almost as much as combining MLP and offset adaptation, except that at very high SPLs, adding offset adaptation enhances the robustness of performance significantly. The results of speech recognition tests showed that MLP generally accounts for the major performance improvement while offset adaptation was essential for the robustness against high sound level.

For experiments under clean condition, the performance gap between speech recognizer and human perception is less than that of the multi-conditional training condition. There is stronger level dependency when the sound intensity is low as compared to multi-conditional training. In multi-conditional training, high SPL deteriorates the speech recognition performance considerably more than low SPL. However, in clean test, low SPL appears to be a more critical factor which deteriotes speech recognition performance significantly.

As in Figure 4.7, the enhanced offset adaptation greatly improves speech recognition at high sound level and slightly degrades the performance at low sound level. MLP, on the other side, always improves the speech recognition performance compared to HTK testbed. The enhanced offset adaptation suppresses the spiking of auditory nerve fibers through the dead-time period. Hence the auditory nerve fibers are less likely to saturate, whereas previously the ANFs saturate very quickly as the sound pressure level increases. On the other side, the suppression of spiking at low sound presssure level leads to too few spikes to generate a meaningful representation of the speech, therefore a deterioration in speech recognition performance.

4.6.2 Speech Recognition using Combined Auditory Nerve Fibers and Octopus Neurons Features

In previous sections, we showed the importance of the proper acoustic modeling and the realistic auditory modeling. In this section, features derived from octopus neuron spike trains were used to augment the rate-place coding strategy from auditory nerve fibers. We chose the system that gave the best performance, i.e., model with enhanced offset adaptation and MLP acoustic modeling. Besides improved speech recognition result, there is another benefit using the MLP testbed: MLPs are very easy to use in a multi-stream approach. It automatically evaluates different features and combine them in an optimal way. In previous work we used KLT to combine features from ANFs and octopus neurons for the HTK testbed. The performance was worse than the current results, thus not discussed here.

Figure 4.9 showed average speech recognition results on vowels at different sound levels. We used features from ANFs and octopus neurons separately, and also the combination of the two. At very low sound levels (25 dB, 35 dB), features from octopus neurons performed much worse than the ANFs, because at low stimuli intensity the spikes of octopus neurons are too sparse to provide meaningful representation of the speech signal. As sound level

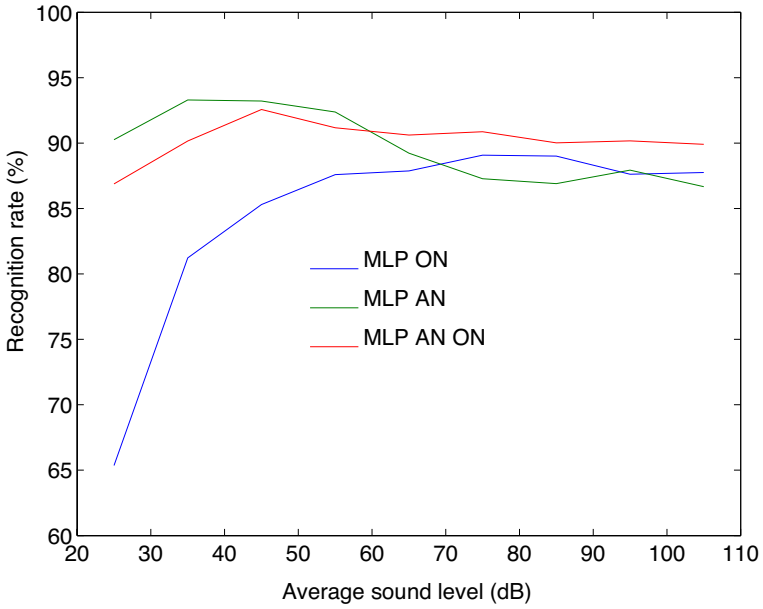

Figure 4.9: Recognition results at different speech level from 15 dB SPL(A) to 105 dB SPL(A) in 10 dB step. The model with enhanced offset adaptation was used and MLP was taken for the acoustic modeling.

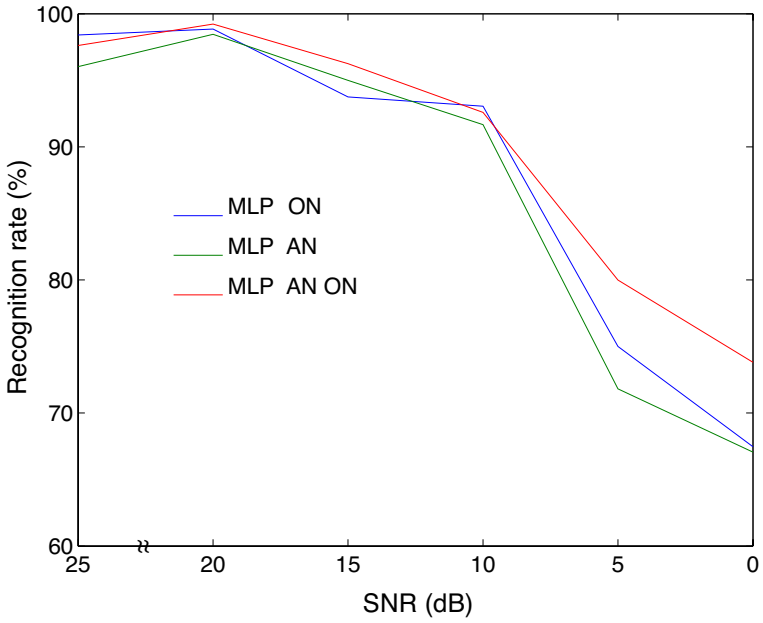

Figure 4.10: Recognition results at different signal to noise ratio. Speech was normalized to 140 dB SPL. The combined features from ANF and octopus neuron achieve better performance than each feature alone.

increased, octopus neuron performance improved considerably and then leveled off. ANF features on the other side, mainly degraded with increasing sound level. The two curves (blue and green) intersected at around 70 dB SPL(A). At high sound levels, octopus neurons gave better or equal performance compared to ANFs, even though octopus neurons had much less spikes.

At low sound levels, combining features from ANF and octopus neuron slightly degrades the performance compared to the result using ANF features only. This is clear because at low sound levels, octopus neurons performed much worse than the ANFs, therefore combining octopus neurons with ANFs actually degraded the performance in comparison to ANFs alone. From sound levels around the conversational strength to even stronger levels, where octopus neurons alone achieved comparable performance to ANFs, combining features from both improved speech recognition results. The performance of combined features was very robust against increasing sound levels.

We took recognition results at 140 dB SPL(A) as an example to show how octopus neurons improve recognition performance at different SNRs. Figure 4.10 showed the corresponding results calculated from different features. Features generated from octopus neuron spike trains outperform ANF features even though that octopus neurons have much sparser spikes. The combined features generated best results at most of the SNR levels. Combing octopus neurons with ANFs improved speech recognition performance most considerably at low SNRs, the relative reduction in WER is 20.0% at 5 dB SNR and 19.5% at 0 dB SNR. This indicates that features derived from octopus neurons improve robustness of speech recognition, thanks to their distinctive temporal processing characteristic – they respond preferably to voiced speech due to their strong phase locking ability to amplitude modulated signals (see Sections 2.3 and 3.3.3).

4.6.3 Comparison with Human Performance

In this section we compared our speech recognition results to human performance, which was represented by the Speech Intelligibility Index (SII, ANSI, 1997). The SII is a measure based on acoustical measurements of speech and noise that is highly correlated with intelligibility of speech. The SII cannot be interpreted as a recognition rate, but is an intermediate result in calculating human speech intelligibility. The actual intelligibility depends on the size and nature of the recognition task, among other factors. To transform the results from SII to recognition rates, we used a transfer function which is also used by Müsch and Buus [2001]. It is based on the transformation between articulation index and an open-set nonsense syllable test, originally suggested by Fletcher and Galt [1950], and is corrected for recognition by chance in the present case of a limited test set. The recognition rate R is described by

$$R = 1 - \left(1 - \frac{1}{M}\right) \cdot 10^{-SII/p}, \tag{4.2}$$

where M is the size of the test set ($M = 26$ in our case) and $p = 0.56$ a fitting constant. Since intelligibility tests with human listeners were out of the scope of this thesis, the

parameter p was estimated from Daly [1987], who reported 1.6% error rate for continuously spoken letters in clean conditions. A parameter value $p = 0.56$ maps a SII of 1 to a recognition rate of 98.4%. Although this relies on a single data point, it should give an idea of human performance. The mapping between SII and intelligibility is kept constant throughout this section. Any changes in intelligibility due to background noise or speech level thus reflects changes in SII values. As the SII considers spectral masking and the human hearing threshold, it replicates the observed decrease in human speech intelligibility at low and high levels [e.g., Beattie and Raffin, 1985, Pollack and Pickett, 1958, Studebaker et al., 1999].

The SII-based recognition scores in this section were calculated using the 1/3-octave band procedure [ANSI, 1997] with speech spectrum levels averaged over all (clean) ISOLET recordings. The level-dependent SII was calculated by varying the level of the speech spectrum. For the estimation of SII in background noise, the spectrum levels of the eight various noise types present in the noisy ISOLET database were calculated. SIIs were calculated for each of the noise types and at all signal-to-noise ratios separately. The SII-based intelligibility predictions presented in Figure 4.7 are averaged over all conditions.

According to the SII prediction, the recognition rate averaged over all SPLs was 91.4%. It is quite clear that there is still a big gap with regard to speech recognition performance between human and the auditory model (average recognition rates were 91.4% and 75.8%, respectively). Regardless of the performance gap, recognition results achieved from the model with enhanced OA (Figure 4.7, and Figure 4.8, solid lines in blue and red) resembled the curve predicted by SII (green line) very well. Recognition performance increased from very low SPL, reached the plateau at around 35 dB to 55 dB and then slowly decreased when SPL further increased. The result showed that the enhanced OA was essential for the auditory model to achieve similar recognition robustness as humans.

4.7 Discussion

The outstanding performance of human speech recognition inspired the effort to model the human auditory pathway. Therefore it is only logical to evaluate the modeling effort with speech recognition tasks. Among others, the generality and its ability to perform complex tasks with various stimuli and many variations in speech make speech recognition a better candidate for evaluating speech coding compare to the discrimination tasks [e.g., Delgutte and Kiang, 1984a, Sachs and Young, 1979, Young and Sachs, 1979, Miller and Sachs, 1983, Sinex and Geisler, 1983, Carney and Geisler, 1986, Deng and Geisler, 1987b, Sachs and Young, 1979, Conley and Keilson, 1995]which tests only well defined properties of speech.

In this chapter, we used speech recognition as a tool to test the performance of the modeled auditory system. In terms of signal processing, speech recognition is used here as a classification task to test the rate coding strategy and the information transmission capability of the auditory system qualitatively. We tested speech recognition performance

for two models – with and without the enhanced offset adaptation – with two different testbeds – HTK testbed using Gaussian mixture model for the acoustic modeling and MLP testbed using multi-layer perceptrons for the acoustic modeling.

The enhanced offset adaptation not only generates realistic auditory nerve outputs per se, but also improves the speech recognition performance. When tested on noisy full alphabet using features derived from auditory nerve fibers, the enhanced offset adaptation improves average speech recognition rate by 12.5% for HTK testbed and 19.6% for MLP testbed. The enhanced OA is vital for octopus neurons to spike properly. Without enhanced OA, octopus neurons stop firing at frequencies above 3 kHz. The improvement brought by the enhanced offset adaptation on vowel recognition with features derived from octopus neurons was more impressive: 37.1% and 38.4% for HTK and MLP, respectively. Given the fact that information was transmitted to the higher processing stage of the auditory pathway only by neurons such as octopus neurons, it is critical to incorporate the enhanced offset adaptation into the auditory system.

Enhanced offset adaptation also shows considerable impact on the robustness of speech recognition at different speech levels. For the original model without the enhanced OA, when the sound level is higher than 45 dB, recognitioin rate using HTK testbed with ANF features dropped very dramatically, at a rate of 9.4% per 10 dB. With the enhanced OA, the recognition degraded only moderately – recognition score dropped only 10.7% in total when SPL increases from 45 dB to 105 dB compared to 56.4% previously. At 105 dB, the relative improvement in speech recognition performance is 95.7%. For speech recognition at different levels under clean condition, the similar impact of the enhanced offset adaptation has been shown.

Human performance in speech recognition represented by the speech intelligence index is considerably better than the performance of the automatic speech recognizer using the auditory features. However, implementing the enhanced OA narrows this performance gap. Despite the difference in performance, the offset adaptation improves the robustness of speech recognition against sound level, which resulted in a curve that is almost parallel to the SII.

Alongside the improvement achieved by using the enhanced OA, adopting MLP testbed also produces extra benifit in terms of speech recognition result. HMM testbed excels in the representation power and the flexibility of training, decoding, and the ability to accommendate different levels of constraints but makes numerous assumptions about speech. We used MLP testbed hoping to avoid the statistical assumptions of the GMM approach and provide useful flexibiliy when working with the model based features. The MLP testbed also allows us to take advantage of the HMM frame work.

The result is promising. Using MLP testbed instead of HTK resulted in major improvements in speech recognition performance for all the auditory based features. This is consistent with past results on MSG auditory features [Sharma et al., 2000, Ellis, b]. However, when using MFCCs on the vowel subset, GMMs outperformed the MLPs. MLPs are also very easy to use in a multi-stream approach. The combined features from ANFs and octopus neurons outperform each single feature. In the future we also hope to ex-

ploit combined features derived from different groups of neurons. Other researchers have found a tandem MLP/GMM approach to be an effective way of incorporating auditory-inspired TRAPs features into a GMM system, while still taking advantage of the GMM system's strengths such as speaker adaptation [Zhu et al., 2004]. We believe researchers working with novel features should consider trying MLPs. The MLP and HTK ISOLET recognition scripts are available online for use with other features.

5 Quantify Speech Information in Spike Trains Using Information Theory

Abstract[1)]

In this paper we use information theory to quantify the information in the output spike trains of modeled cochlear nucleus onset neurons (ONs). Onset neurons are known for their precise temporal processing, and they code the periodicity of voiced speech with high fidelity. We conclude that the initial maximum information transmission rate for a single neuron is close to 1050 bits/s, which corresponds to approximately 5.8 bits per spike. For quasi-periodic signals like voiced speech, the transmitted information saturates as word duration increases. In general, approximately 90% of the available information from the spike trains was transmitted within 45 ms. Information of speech concentrates at formant frequency regions. The efficiency of neural coding lies above 60% up to the highest temporal resolution we investigated (20 μs). The increase in transmitted information to that precision indicates that these neurons are able to code information with extremely high fidelity, which is required for sound localization. On the other hand, only 15% of the information was captured with a temporal resolution of 10 ms. As the temporal resolution of most speech recognition systems is limited to this value, this massive information loss might explain the lack of noise robustness of these systems.

[1)] A modified version of this chapter has been submitted to the Journal of Computational Neuroscience. Some of the similar results were also published in the proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2006 [Wang et al., 2006].

5.1 Introduction

The outstanding performance of the human auditory system, especially in noisy environments, motivates the investigation of auditory principles for robust coding and recognition of speech. Our auditory system performs a spectral decomposition of acoustic stimuli, and at the same time preserves temporal information by the virtue of precisely timed nerve-action potentials. In this paper we investigate the coding strategy of the initial neural processing stages. The following problems are of particular interest for us: How is speech encoded into neural spike trains? How is information distributed across frequency bands? What is the temporal resolution of the neural code, and in which time scales is most information represented? And furthermore, how efficient and how robust is the coding in noise?

Neural coding has been classified into two major types: rate codes which consider only the average rate of spike discharge, and temporal codes, in which the timing of individual discharges matters [Theunissen and Miller, 1995]. Debates are still ongoing over the effectiveness of one code over another [Holmberg et al., 2007] and most automatic speech recognition algorithms use short-term spectra and neglect – or better discard – fine-grained temporal information. Early theories of neural information processing, dating back to the very influential observations by von Helmholtz [von Helmholtz, 1863], generally assumed that all the information carried by neural spike trains is contained in the firing rate [Adrian, 1928]. However, the accuracy of individual spike timing represents a second, if not the most important, mechanism of information transmission along the neural pathway [Rieke et al., 1997]. Along these lines, we have shown that the rate-place code alone is not sufficient for the robust discrimination of noisy speech at high sound levels and highlighted the importance of temporal coding at least under these adverse conditions [Holmberg et al., 2007].

In Chapter 4 we have utilized speech recognition to evaluate the neural code and have shown that the spike trains of auditory nerve fibers together with spikes from onset neurons provide an effective representation of the speech signal. However, the recognition performance was not particularly better than MFCC feature extraction, which still can't get even close to human performance (SII). It was not very straight forward to draw conclusions about neuronal coding based on the speech recognition results. Besides, only the rate-code principle was examined. As we used a 25 ms Hanning window forwarded in 10 ms steps, information encoded in a temporal scale finer than 10 ms was discarded. It's hard to fully exploit the coding capacity of the neuronal spike trains just using speech recognition. Therefore in this chapter, we utilized information processing to evaluate the neuronal encoding strategy.

In our calculations, we therefore represent spike trains with the precise timing of each individual spike. This representation makes no assumptions on the coding strategy, it thus evaluates all the information that is carried by the neural outputs. We generated spike trains with our previously developed auditory model of the human inner ear [Holmberg and Hemmert, 2004, Holmberg, 2007] with an extension of enhanced offset adaptation

[Wang and Hemmert, 2007] and applied information theory to find answers to the above questions. We evaluated the robustness of speech coding for Type-II onset neurons (especially the spherical bushy cells) located in the cochlear nucleus (CN). The CN is the first neural processing stage after the inner ear. ONs respond with great precision to signal onsets and they extract the periodicity of voiced speech with high fidelity [Hemmert et al., 2005].

5.2 Modeling Spike Trains of Onset Neurons

We used the peripheral hearing model that was introduced previously. The model includes the frequency response functions of outer- and middle ear, a model of inner ear hydrodynamics followed by a compression stage, and sensory cells (see Figure 2.1). In summary, the hydrodynamic model effectively acts as a filter bank that spectrally decomposes acoustic stimuli. The compression stage models the effects of the so-called "cochlear amplifier". It achieves up to fourth-root compression of the dynamic range. The absolute compression is more than 60 dB. It also achieves the high spectral resolution found in humans at low sound levels. Large compression is crucial for sound coding by the inner-hair cells, since they have a dynamic range of only 40 dB. Auditory nerve fibers innervate the sensory cells and encode the stimuli into nerve action potentials, i.e. spike trains. The generation of a spike by an ANF is implemented using the pool model, which we further extended to reproduce an effect termed offset adaptation. Our model generates realistic ANF responses which reproduces recent psychoacoustic measures of frequency selectivity and compression [Holmberg and Hemmert, 2004]. It replicates the bandwidths of human threshold tuning curves [Shera et al., 2002] and latest measurements of dynamic range compression [Lopez-Poveda et al., 2003] with great precision.

We modeled Type-II onset neurons [Rothman and Manis, 2003a] located in the auditory brainstem (ventral cochlear nucleus VCN) and connected them to 60 ANFs (all with the same characteristic frequency) from our inner ear model (compare Figure 2.1).

Figure 5.1 presents the responses derived from the utterance /*ei*/ (female speaker, ISOLET [Cole et al., 1990] fcmc0-A1, scaled to a sound level of 75 dB(A)). ISOLET recordings are band-limited to 8 kHz. The responses of the ANF and ON along the length of the inner ear are plotted in Figure 5.1. Note that the inner ear performs a spectral decomposition with approximately logarithmic frequency resolution. ANF spike trains code both spectral- and temporal features of sounds. Frequency regions with high energy – like the formants – are coded with higher spike rates. The temporal fine structure of the sound signal is preserved in the precise spike timing. The increasing delay of the neural responses towards lower frequencies is caused by the propagation of the traveling-wave from the base to the apex of the inner ear. ON enhance the periodicity of voiced speech; especially in the frequency region above 1kHz, they lock on the time structure of the acoustic signal and reliably extract the pitch frequency [Hemmert et al., 2005]). At lower frequencies, they are entrained directly with the characteristic frequency of the auditory nerves, to which they are connected.

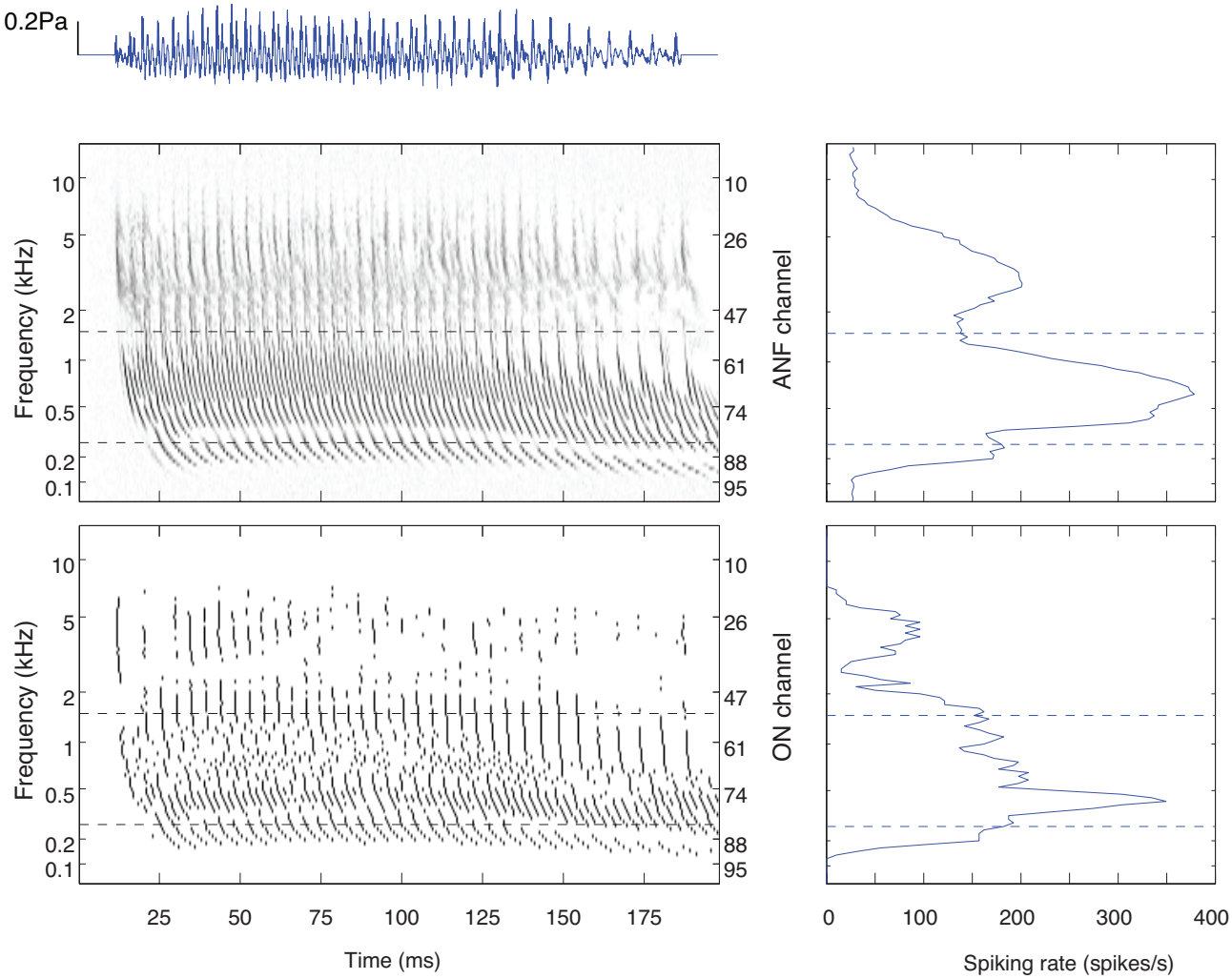

Figure 5.1: Coding of speech (utterance /*ei*/, female speaker, 75 dB(A)) into trains of nerve-action potentials of the auditory nerve fibers (upper left pannel) and onset neurons (lower left panel).

5.3 Information Calculation

Since sensory systems are analog to communication channels [Attneave, 1954], a natural approach to analyze neural coding is to use mutual information. Let S denote the input stimulus and R the neuronal response. Based on Shannon's information theory [Shannon, 1948], given the neuronal spike trains, the information $I(R;S)$ conveyed about the input stimulus can be calculated by

$$I(R;S) = H(R) - H(R|S) \tag{5.1}$$

where the entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x} p(x) \log_2 p(x) \tag{5.2}$$

and the conditional entropy $H(X|Y)$ of X given Y by

$$H(X|Y) = -\sum_{y} p(y) H(X|Y = y) \tag{5.3}$$

We followed Strong's direct method [Strong et al., 1998] to compute the mutual information and the total response entropy of individual neurons in the auditory pathway. This approach makes no assumptions how the neurons code input stimuli and thus preserves all the information which is contained in the spike trains.

Time (ms)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...
	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	...
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	...
	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
Subset 1	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	...
	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	...
	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	...
	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	...
	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	...
	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
Subset 2	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	...
	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	...
	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	...
	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
.	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	...
.	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
.	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	...
.	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	...
.	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
.	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	...
	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	...
	first word																
		second word								.....				11 *th* word			

Table 5.1: **Binary representation of onset spikes.** This table is based on realistic output of a neuron with time quantization of 1ms. The rows correspond to different trials.

We down-sampled the (quasi-) continuous response of the onset neurons into time-bins with appropriate length. We will refer to this down-sampling operation later as a "binning" process (please see Nelken et al. [2005] for a description of different estimation methods of mutual information, including a binless approach [Victor, 2002]). We then represented these discrete bins with binary letters. Depending on whether there was an output spike within the bin or not, the letter evaluated to "1" or "0", respectively. Hence we obtained a discrete version $T(R)$ of the neural response. The discrete spike train $T(R)$ of a given neuron was further translated into a sequence of words along the time axis. Each word W consisted of L letters, while each letter covered a time bin of ΔT ms, which was the temporal resolution of the binning process. Table 5.1 shows how the spike trains of an onset neuron was represented by binary digits and further grouped to generate words. For a given spike train, we computed the histogram of occurrences for each possible word (duration: $L\Delta T$ ms), which was then used to calculate the probability of the word, i.e., $p(W(L,\Delta T))$. Then, the entropy per unit time, or equivalently, entropy rate, was computed:

$$H_r(L,\Delta T) = -\frac{1}{L\Delta T}\sum_W p(W(L,\Delta T))\log_2 p(W(L,\Delta T)) \tag{5.4}$$

If each bin of the output spike train was independent, then the calculation with any word length L would give us the same entropy. However, there are correlations both due to the correlations in the input stimulus and due to the processing within the auditory pathway, e.g. refraction. To evaluate these effects, we studied the how entropy is influenced by word length.

The noise entropy (or equivalently, conditional entropy) represents variations between different responses of a neuron when the stimulus is repeated. It was computed the following way: for M repetitions of the stimulus we get M trials of spiking responses. The occurrences of words $W(L,\Delta T)$ at a particular time t_i leads to the conditional probability of occurrence $p(W(L,\Delta T)|t_i)$. Entropy for the distribution at time t_i was computed and then averaged over time to obtain the noise entropy rate,

$$H_r^{noise}(L,\Delta T) = \left\langle \frac{-\sum_W p(W(L,\Delta T|t_i))\log_2 p(W(L,\Delta T|t_i))}{L\Delta T} \right\rangle_n \tag{5.5}$$

where $\langle \ldots \rangle_n$ denotes the average over the whole sampling time from t_0 to t_n.

We also used information theory to estimate the temporal precision of the action potentials. Information was calculated using different temporal resolutions for the binning process. According to the data processing inequality[2)], the input stimulus S, the response R of an onset neuron and its discrete representation forms a Markov chain $S \to R \to T(R)$, therefore the information calculated from $T(R)$ is always less than the real amount of

[2)] Data processing inequality: If $X \to Y \to Z$ form a Markov chain which means X and Z are independent given Y, then $I(X;Y) \geq I(X;Z)$ [Cover and Thomas, 1991b].

information. The loss of information decreases when we use a finer time resolution for the binning process. It goes to zero when the binning resolution is higher than the actual temporal resolution of the action potentials. In other words, if the binning resolution is already higher than the precision of the generated spikes, then using an even finer resolution will lead to no further increase in information transmission. Therefore, we searched for the point where information saturates when increasing the binning resolution [Borst and Theunissen, 1999].

In our experiments, we presented speech signals to our auditory model and evaluated the distribution of firing patterns of the ONs. We used responses from ON because they integrate over many ANF inputs. Therefore, ON responses are much more reliable compared to ANFs, which immensly reduces the number of occurring spike patterns. For this reason, information calculations using ON spike trains converge faster than for ANFs. We used the type-II bushy cells to generate the output spike trains. Bushy cells behave very similarly to octopus cells, except that bushy cells do not have the hyperpolarization-activated cation current I_h. Bushy cells have a lower spiking rate than the octopus cells, therefore they are easier to calculate using information theory[3]. In Chapter 6, we will also show some results from octopus neurons.

5.4 Results

5.4.1 Feasibility of Robust Estimation

Estimating entropies from empirical distributions requires large amount of data. According to Rieke et al., we can relatively easily estimate the upper bound of entropy of a certain spike train [Rieke et al., 1997].Imagine a spike train with an average spiking rate $\overline{r}$ and a total duration of t observed with a temproal resolution of ΔT, where ΔT was small enough so that in each bin there is either one spike or none. The upper bound of entropy can be estimated assuming that spikes in different bins are uncorrelated,

$$H = -\frac{t}{\Delta T}(\overline{r}\Delta T \log_2(\overline{r}\Delta T) + (1 - \overline{r}\Delta T)\log_2(1 - \overline{r}\Delta T)) \tag{5.6}$$

In our case, the onset neuron had an average spiking rate of about 156 spikes/s at a characteristic frequency of 1.5 kHz[4]. We sampled the spike train with a resolution of $\Delta T = 4$ ms, which is small enough to guarantee that there is at most one spike in each bin.

[3] We did the most calculation for both bushy cells and octopus neurons, therefore the conclusions in this chapter applies for both types of cells. However, for octopus cells, the information calculation requires more trials to see a saturation at long word lengths. We didn't calculate how the pink noise affects the information distribution at different channels either due to computational cost.

[4] In this frequency channel the firing rate was relatively low. We also show resuls from the 260 Hz channel, which is located in the region of the first formant, where ONs in our model fired with the higest rate.

The longest word duration on which we calculated entropy was 72 ms. Correspondingly, the maximum entropy can be estimated to about 17.2 bits (Equation 5.6). Theoretically, it would require at least 3 hours stimulation to cover all the instances if non-overlapping words were used. For a word duration of 100 ms, the required theoretical stimulation period would increase drastically to 435 hours. Therefore, for long word durations it becomes increasingly hard to achieve robust estimations.

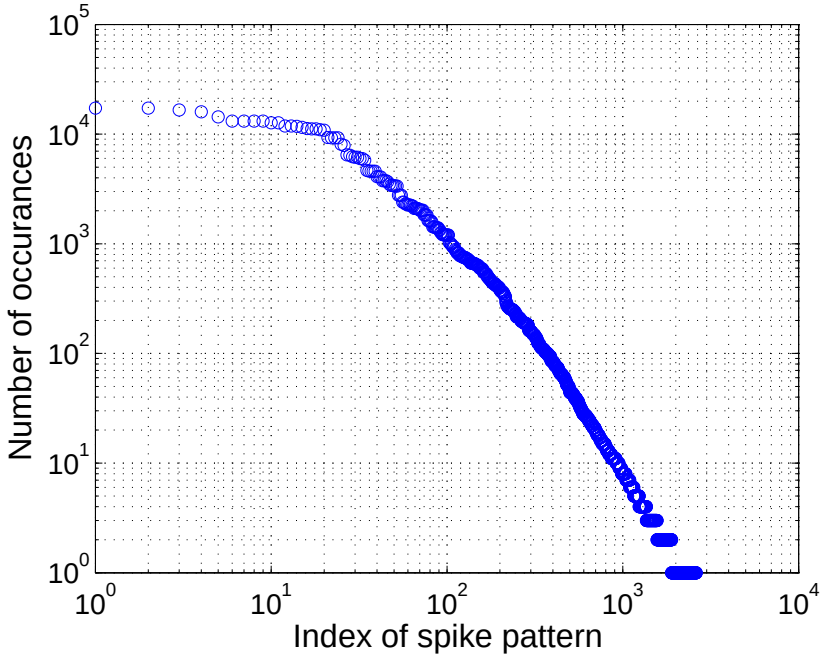

Figure 5.2: The distribution of output spike train patterns. The outputs correspond to about 1 hour neural recordings. Spike trains are grouped into words of 72 ms duration with 4 ms temporal resolution. Around 2617 different patterns in total were observed. The maximum number of occurrences for a single pattern amounts to 17,276. The observed distribution is very peaky — 1% of the patterns which occur most frequently contribute to 49% of the total occurrences, whereas 27.7% of the patterns occur only once. For better visualization, the figure was plotted on double-logarithmic axes.

Despite this pessimistic outlook, we found that a robust estimation is feasible. First, the patterns of ON spikes which actually occur are much less than the theoretic estimation, about 2617 in our calculation compared to 149,835 resulted from the theoretic estimation. Second, the distribution of the possible instances is extremely peaky (see Figure 5.2), which leads to a small estimated entropy of about 7.4 bits. This entropy is a maximum likelihood estimation of entropy (H_{ML}), since the maximum likelihood estimate of the probabilities is given by the frequencies of occurence. According to [Nemenman et al., 2004], the number of samples used in our estimation ($\sim 2^{19.3}$) is much larger than $2^{H_{ML}}$ ($H_{ML} = 7.4$), indicating that the direct estimation has already converged to its asymptotic value. In fact, since we used overlapping words in our calculation, the total duration of

the signals corresponded to about 35 hours, which was much longer than the 3 hours required to cover all the possible instances. It thus indicated that a robust estimation is possible. Besides these two facts, we also applied various techniques to compensate the possible under-estimation problem in the following parts of this paper. And finally, our model enables us to achieve even larger amounts of experimental data as long as the computational power is available.

5.4.2 Information Conveyed by Spike Trains

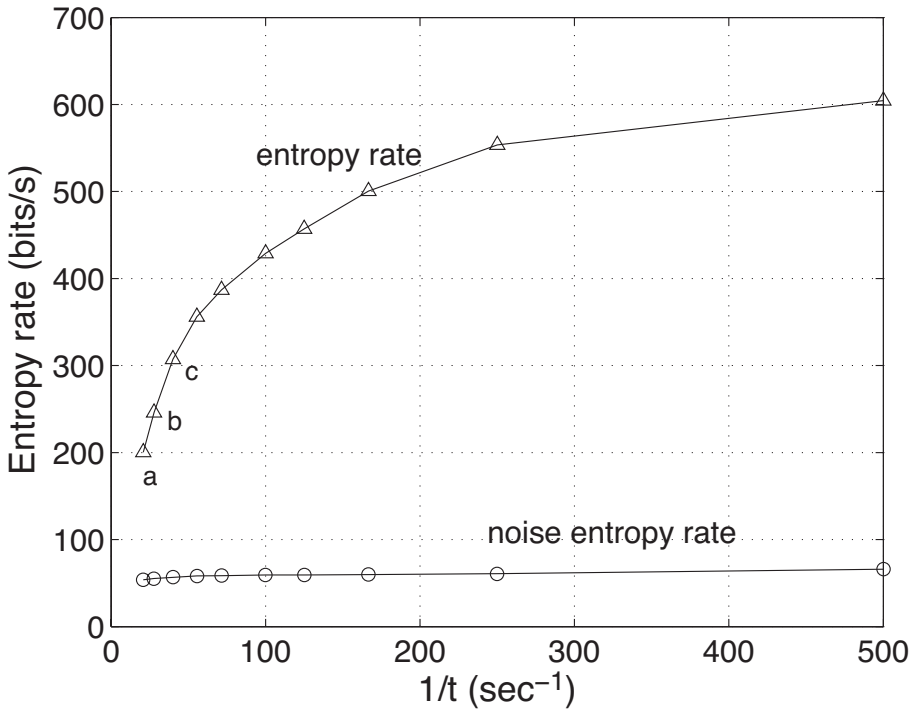

Figure 5.3: Relationship of total- and noise entropy rates on the reciprocal of word duration. Spike trains were taken from an ON with a characteristic frequency of 1.5 kHz. The binning resolution was 1 ms. Results were calculated from 20,000 repeated experiments with the same input data.

Figure 5.3 shows total entropy rate and noise entropy rate calculated by the algorithm described in Chapter 5.3. The relationship between noise entropy rate and the reciprocal of word duration can be well described by the following linear function,

$$H_r^{noise}(L, \Delta T) = \frac{H_{noise}(L, \Delta T)}{L \cdot \Delta T} = r + \frac{m}{L \cdot \Delta T} \tag{5.7}$$

which in turn gives us the following relationship between noise entropy and word duration,

$$H_{noise}(L, \Delta T) = r \cdot L \cdot \Delta T + m \tag{5.8}$$

where r denotes the rate of noise entropy in bits per time unit at infinite word duration and m a constant, which corresponds to the slope of the noise entropy rate in Figure 5.3. Asymptotically, $\lim_{L\to 0} H_{noise}(L, \Delta T) = 0$ leads to $m = 0$, which is in accordance with our calculations, where noise entropy rate is almost constant with a standard deviation of only about 5.1%. Still, in the following numerical estimations, we kept the m value from the linear extrapolation, even though it was very close to 0.

The dependency between total entropy and word length is not as simple. In Figure 5.3, the total entropy rate as a function of the reciprocal word duration is plotted. Our results are similar to previous data (e.g. [Strong et al., 1998, Yu et al., 2004]). The entropy rate decreases monotonically with increasing word duration. The decrease is moderate and almost linear for short word durations. When the total entropy rate as a function of the reciprocal word duration is plotted, usually the linear part of this data is extrapolated to infinite word durations [Strong et al., 1998]). However, our data shows a hyperbolic decrease of the entropy rate. This decrease is usually declared as a result of an insufficient number of samples ([Strong et al., 1998])), due to the limited recording times available from experimental recordings. However, in our case, we were able to overcome this problem with a large number of computational samples and we will show that this hyperbolic decrease has other systematic reasons.

First we show the reliability and convergence of our data. Numerical convergence becomes an increasingly severe issue for long word lengths, because the number of possible words increases exponentially. We have therefore investigated the data points with the three longest words, corresponding to a word duration of 25 ms, 36 ms and 48 ms. To investigate numerical convergence, we plotted the mean value and the standard deviation of the total entropy rate as a function of the number of trials using a logarithmical abscissa (Figure 5.4). For word durations of 25 ms and 36 ms, the value of both mean entropy rate and standard deviation clearly indicated a robust estimation for a large number of samples; the mean of the estimated entropy rate changed by less than 1% when the number of trials exceeded 100 trials.

For the word duration of 48 ms, the entropy rate increased systematically as the number of trials increased. The entropy rate did not fully converge even when we used the largest trial number. However, the increment leveled off and the plot clearly showed that the entropy rate was almost converged. We may expect that the asymptotic value would (almost) certainly not exceed 202 bits/s.

In summary, we achieved a reliable estimation of the entropy rates even at the longest word lengths we used in this paper. Therefore we have to conclude that a linear extrapolation of the total entropy rate to infinite word lengths is not appropriate in our case. Instead, the total entropy rate decreases systematically and hyperbolically for increasing word durations.

Since we have excluded the under-sampling problem as the reason of the drop of total entropy rates with word duration, we have to describe the relationship between total entropy rate and duration in a different way, rather than a simple linear one. Instead of fitting entropy rate and the reciprocal of duration, we now look at entropy as a function

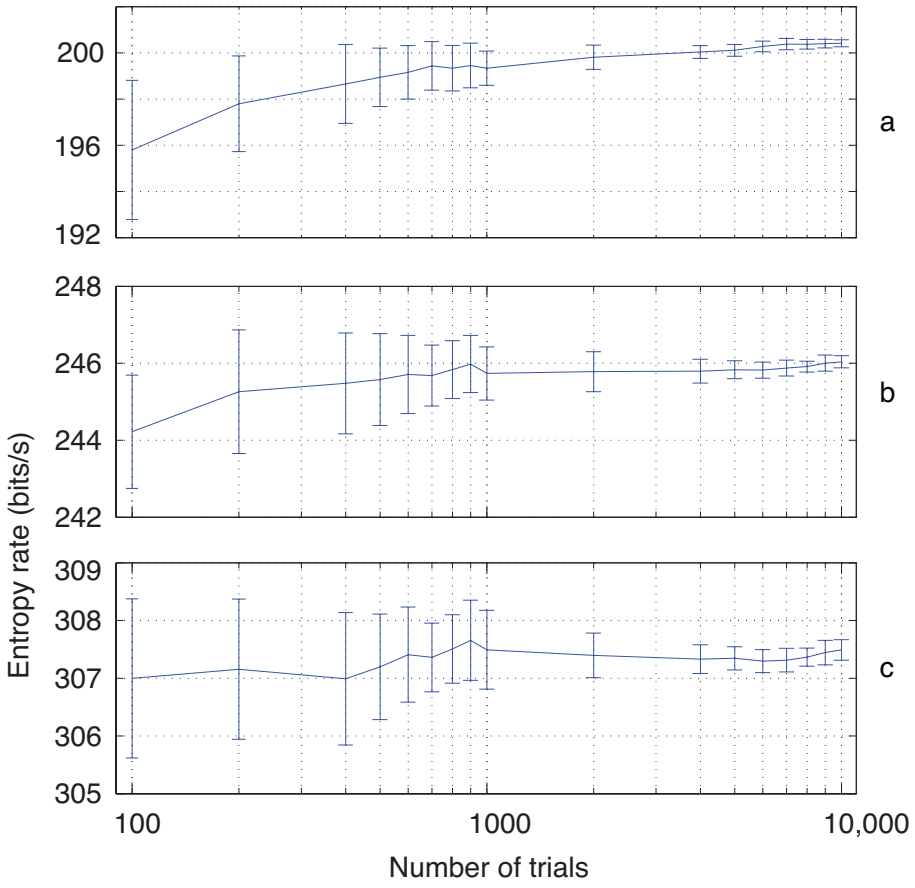

Figure 5.4: Entropy rates calculated from different numbers of samples. Panel a,b and c correspond to the three datapoints with the longest word lengh (word length 48, 36 and 25 bins, resolution 1 ms) from Figure 5.3, where the under-sampling problem becomes most severe. Mean values from 8 identical calculations are plotted. The error bars show one standard deviation above and below the average.

of word duration, which is easier to understand in this respect. As shown in Figure 5.5, noise entropy is roughly linear as a function of word duration (Equation 5.8). For the total entropy, there is an initial steep increase which then levels off to the same steepness as for noise entropy. This can be seen more clearly in transmitted information trace (which is the difference between total entropy and noise entropy) shown in the right panel of Figure 5.5. Transmitted information increases steeply for small word durations and saturates at longer word durations. We applied different possible functions to fit the data and found an exponential function best described the experimental data. We will illustrate in the discussion that this function is also analytically meaningful. Total entropy can be described by

$$H(L, \Delta T) = r \cdot L \cdot \Delta T - c \cdot \exp(-L\Delta T/\tau) + d \tag{5.9}$$

where r takes the same value as in Equation 5.8. c, τ and d are coefficients to be estimated

by fitting the experimental data. Asymptotically $\lim_{L\to 0} H(L, \Delta T) = 0$ gives us $d = c$. Therefore the information carried by the spike trains can be calculated:[5)],

$$\begin{aligned} I(L, \Delta T) &= H(L, \Delta T) - H_{noise}(L, \Delta T) \\ &= c(1 - \exp(-L\Delta T/\tau)) \end{aligned} \tag{5.10}$$

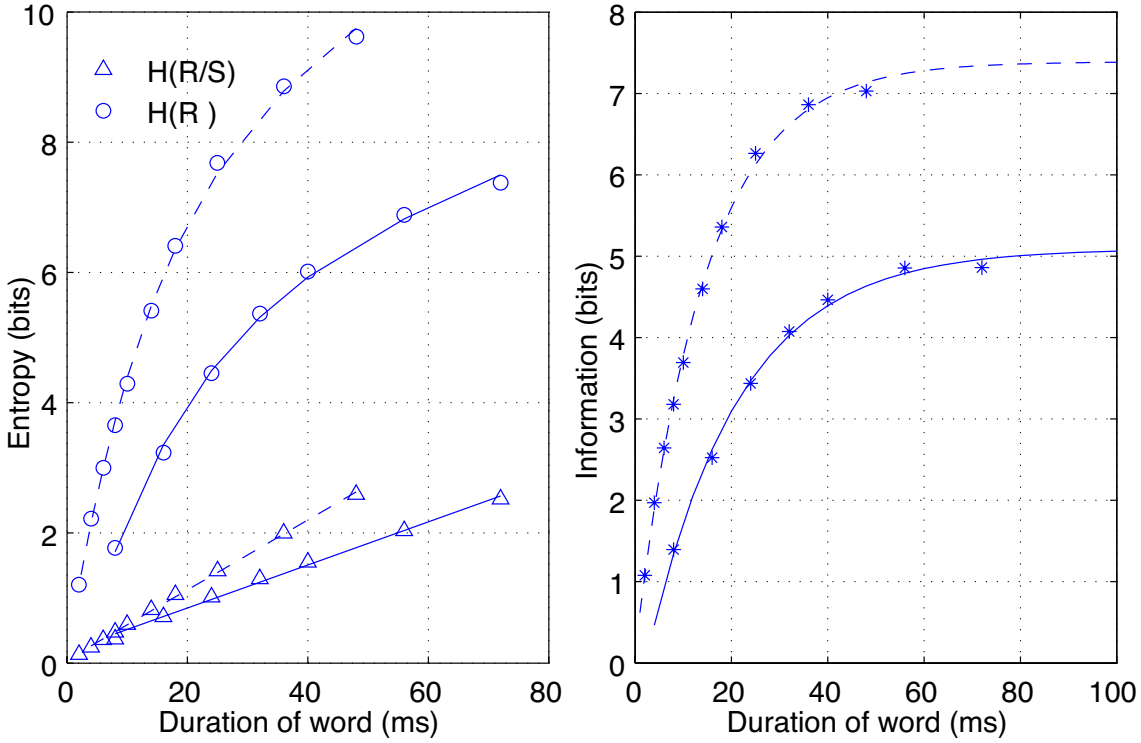

Figure 5.5: Dependency of total entropy, conditional entropy (left pannel), and information (right panel) on word duration. Round circles and triangles refer to calculation from the experiments. Smooth lines are fits to the data. Two sets of experiments with binning resolutions of 4 ms (solid line) and 1 ms (dotted line) are performed. With a coarser binning resolution (4 ms), less information is transmitted but we are able to cover longer duration with the same word length L and thus see the saturation of information more clearly. At all conditions, the experimental data is fitted well by the functions.

Our fit predicts that the information conveyed by the output spike trains of the onset neuron has an initial steep increase and then levels off asymptotically to a constant amount. Information conveyed by the onset neuron saturates as word duration increases. 90% of the available information is transmitted within 45 ms. When we compare the two curves calculated for 1 ms and 4 ms resolution, it is also clear that for a coarser resolution, we miss information which is preserved when using finer temporal resolution. We will address this point in the next chapter.

[5)] The numerically minor constant part is neglected in our formula, though we keep it in our computations.

5.4.3 Temporal Resolution of Spike Trains

We calculate the information carried by onset neuron spike trains using different binning resolutions. It is obvious that at different resolutions, words with the same length L cover different durations ($L\Delta T$) of the signal. In order to make a comparison between results meaningful, we use words of 4 ms duration[6)] for all the calculations and compare the information rate for different resolutions.

Figure 5.6 shows how information rate depends on the resolution that is used to represent the spike trains. As the resolution increases, information rate increases monotonically, in consistency with the Data Processing Theory. The highest temporal resolution we could investigate in our study is limited by the sampling frequency (48 kHz), which is equivalent to a temporal resolution of about 20 μs. The increasing information rates up to this limit illustrates that still some information is coded in the spike trains of onset neurons even at the highest temporal precision accessible to our study. If we compare the highest information rate to the information at a resolution of 10 ms, which is used for most automatic speech recognition systems, we see that only about 15% of the available information is processed. At 0.25 ms resolution, the information being transmitted corresponds to about 80% of the overall information. Therefore in the range from 0.25 ms to 10 ms, about 65% is coded, or equivalently, 65% of the information lies in frequency regions between 100 Hz to 4 kHz [Borst, 2003].

We further investigated the effect of different temporal binning of the neural spike trains by looking into the efficiency of the neural coding. As the information that the spike train provides about the stimulus is simply the difference between the total entropy and noise entropy (see Equations 5.1,5.10), and noise entropy is positive (semi)definite, therefore the total entropy $H(L, \Delta T)$ sets the capacity for transmitting information. The efficiency with which this capacity is used is defined by [Strong et al., 1998]

$$\epsilon = I(L, \Delta T)/H(L, \Delta T) \tag{5.11}$$

The question of how important spike timing is can be reformulated to the question of whether this efficiency is high at small ΔT [Strong et al., 1998].

Figure 5.7 shows the efficiency of information transmission using different binning resolutions on the spike trains. At most resolutions for both characteristic frequencies, the efficiency of neural coding remains above 60%, in accordance to what other researchers have shown [Borst and Theunissen, 1999]. At resolutions between 0.25 ms and 1.5 ms, neurons with both characteristic frequencies show very high efficiencies in coding. Increasing the resolution above 0.25 ms or decreasing it below 1.5 ms degrades the efficiency.

As we have mentioned before, we can decide whether spike timing is important from evaluating the efficiency, i.e., if efficiency remains high at fine temporal resolution, then spiking timing at this scale is still important. Figure 5.7 shows that even at the highest

[6)] The duration is limited by the word length at highest resolution, which is constrained by the computational expense. At 1/48 ms resolution, 4 ms corresponds to word length of 192.

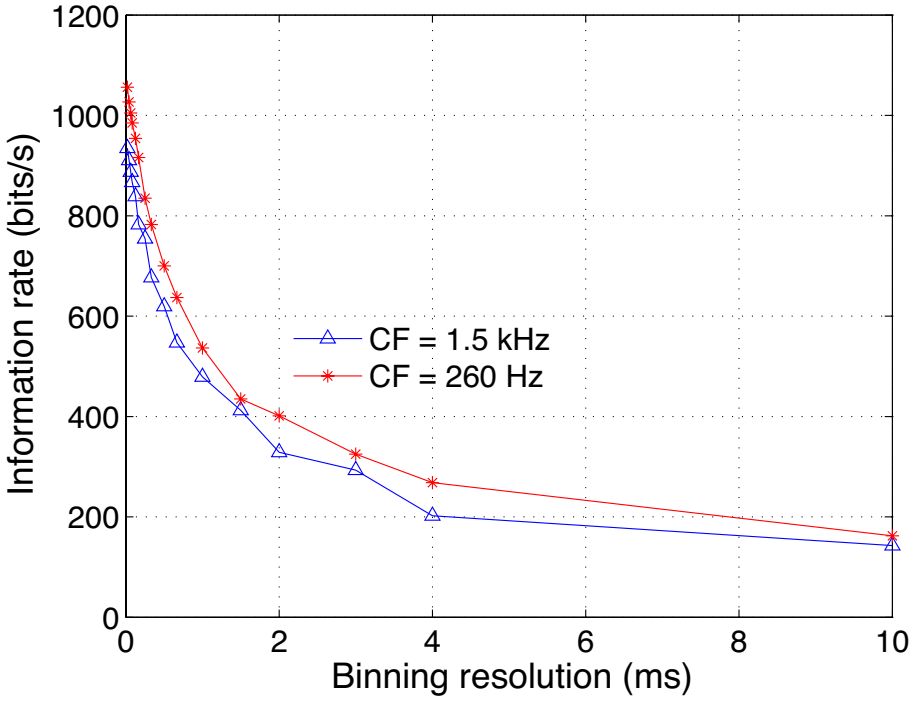

Figure 5.6: Information rate calculated for different binning resolutions. For each calculation, the word duration is set to 4 ms. At 10 ms resolution, we used the fitted curve to calculate the information rate for 4 ms word duration. Input Stimuli and output spike trains are the same as in the previous tests. The abscissa shows the resolution for our binning process. Results of the 1.5 kHz frequency channel and the 260 Hz frequency channel are plotted.

resolution that we can simulate, neural spike trains achieve fairly high efficiency of around 65%. At 0.25 ms resolution where approximately 80% of the information is transmitted, the efficiency for both channels is as high as 88%. This indicates that Onset Neurons do exploit a very high temporal resolution to code most of the information with high efficiency. It is yet to be found out what kind of specific information the Onset Neurons are trying to code and whether this information is also relevant for robust speech coding or other cognitive tasks.

At all resolutions, the 260 Hz channel achieves higher efficiencies than the 1.5 kHz channel, which is attributed to the reliable locking to the characteristic frequency.

5.4.4 Information Distribution over Frequency Channels

Frequency decomposition of speech signals is essential for speech perception. In this chapter we investigate how information is decomposed and distributed over frequency channels and how robust onset neuron spike trains are in noisy conditions.

Figure 5.8 shows that information coded by the onset neurons is distributed over a wide

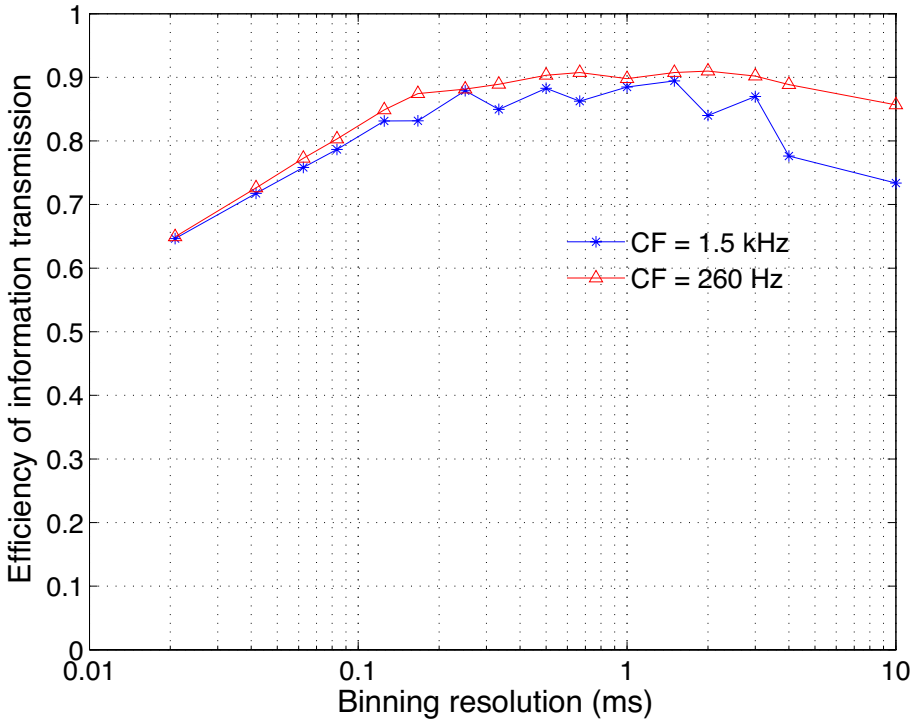

Figure 5.7: Efficiency of information transmission ϵ computed at various temporal resolutions, from $\Delta T = 10\,ms$ to $\Delta T = 1/48\,ms$. Note the original spike trains coming out from our auditory model is sampled at 48 kHz, therefore 1/48 ms is the highest resolution we can achieve.

frequency range. In clean conditions (30 dB SNR), the information rate reached high values in the regions where ONs spiked most frequently [Borst and Haag, 2001]. The largest portion of information concentrated at low frequencies from about 100 Hz to 600 Hz, where the spikes precisely phase-locked to the pitch frequency or its higher harmonics. Also in the 1-2 kHz and 3-5 kHz regions, the information rate reached comparable high values. In these regions, ONs locked reliably on the pitch frequency of the voiced speech signal.

The spiking rates of ONs reached their maxima at CFs around 440 kHz, where the ONs phase locked to the second harmonic of the pitch frequency. The firing rate at the 440 kHz CF was almost twice as high compared to the rate at pitch frequency but the information rate was only slightly higher.

We investigated the robustness of neuronal coding by adding pink noise ar various SNR ratios. For comparison, we kept the speech level constant. Therefore the total level of speech with noise increased with falling SNR. If one would just add a pink noise to a speech signal, the information algorithm would not be able to distinguish between speech and noise information; it would just estimate the information of the input signal, e.g. speech and noise. In order to investigate how noise degrades speech information, we recalculated the pink noise for every trial, added it to the speech signal, presented it to the inner ear model and analyzed the spike trains. Using this procedure, we only estimated speech

information, as the information of the pink noise was different from trial to trial.

Adding pink noise to the signal gradually deteriorated the information transmitted by the ONs. For a SNR of 0 dB, pink noise corrupted about 73% of the information at the 260 Hz location and 85% at the 1.5 kHz location.

5.5 Conclusions and Discussion

Onset neurons located in the cochlear nucleus are known for their distinct temporal processing capabilities. In this paper, we analyzed their performance from the aspect of information transmission. We used speech data, fed it into our computational inner ear model which generated realistic auditory nerve responses. We used these to drive our onset neuron model. We analyzed the neural spike trains using the direct method based on Shannon's information theory to calculate noise entropy and transmitted entropy of individual neurons with different characteristic frequencies. The direct method does not make any assumptions on the neural code and therefore preserves all the information present in the spike trains.

We found that the information rate in our case is not constant for increasing word lengths. Instead, transmitted information for voiced speech sounds saturated. 90% of the information could be transmitted within about 45 ms in general. The time constant of the exponential saturation curve depended on signal level; for louder signals the information is transmitted faster (data not shown).

In the following paragraph we will analyze this behaviour in detail and provide an explanation. Noise entropy rate was indeed almost constant with word duration. Therefore noise entropy increased linearly with word duration (see Equation 5.8), as expected. The speech sound we input to our model is a quasi-periodic signal, which repeats itself – at least approximately – every pitch period. Onset neurons fire reliably at each pitch period of the input signal. Therefore, for a given stimulus, the uncertainty of the output spikes lies mostly in the jitter of each individual spike timing (firing rate is relatively constant for different trials even within a small time window), which is only dominated by the same stochastic property of our auditory processing. In this case, conditional entropy can be viewed as a sequence of variables which are almost independent and identical. According to information theory, the joint entropy of IID (identically and independently distributed) variables can be calculated as the sum of all entropies, namely,

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) = nH(X) \tag{5.12}$$

In fact, this is equivalent to Equation 5.8, which fits our experimental data nicely, if we neglect the small constant m.

In the next step, we demonstrated that the decrease of the total entropy rate with increasing word length (Figure 5.3) cannot be interpreted as an under-sampling problem in

our case. In our analysis of the numerical convergence we plotted the estimated entropy as a function of trial number on a logarithmc scale over two orders of magnitude. This is computationally expensive but proves that our estimation is robust and the fast drop must arise from another mechanism. Equation 5.9 and Equation 5.10 are derived numerically based on the following two observations: First, total entropy rate decreases as duration increases. As mentioned in Chapter 5.3, if there were no correlations within the output spike trains, estimating the entropy for any word duration would give us the true entropy rate. But there are correlations both due to the periodic character of voiced speech signal we investigated and correlations introduced by neuronal processing. These correlations will be taken into account by increasing word duration. Therefore, the total entropy rate decreases with word duration. Second, if word duration is long enough, total entropy and noise entropy increase with approximately the same rate. We believe that this is also caused by the quasi-periodic property of our input signal and the specific behavior of onset neurons, i.e., their firing depends only on the input stimulus of about several milliseconds before the spike [Hemmert et al., 2005, Svirskis et al., 2004]. The derived equations based on these observations fit the experimental data precisely. The information in a coded vowel therefore does not increase with word duration but saturates within some tens of milliseconds. This is again very plausible: the quasi-periodic structure of a vowel repeats itself and does not add additional information. In other words, we transmit the same anount of information, no matter how long we pronounce a vowel.

If we now restrict our interest to the maximum initial information transmission rate, we found values close to 1050 bits/s, which corresponds to the comparatively high information rate of 5.8 bits/spike. We also studied the robustness of neuronal coding to noise. For pink noise with 0 dB SNR, the information content of voiced speech decreased significantly compared to signal at high SNR (30 dB). The decrements for the 1.5 kHz channel and the 260 Hz channel are 85% and 73% respectively. Information concentrates at formant regions of the input speech. The distribution pattern of information is well preserved at high resolutions (10 dB to 30 dB). At an SNR of 0 dB, the pattern was considerably deteriorated.

Our results also show that the temporal resolution of onset neurons is very high from an information transmission point of view. We found that ON code temporal information up to the highest resolution we could investigate: 20 μs. This high resolution is required for sound localization in the horizontal plane, where ON are believed to play a major role.

With a temporal resolution of 0.25 ms, still about 80% of the information is transmitted. Pichora-Fuller et al. showed in their study [Pichora-Fuller et al., 2007] that destroying temporal information at a precision higher than 0.25 ms in the frequency region below 1.2 kHz already degrades speech intellegibility of young healthy listeners in noise. Their performance then becomes identical to the results obtained from older adult listeners with good audiograms when there is no temporal jitter in the signal. This finding indicates that precise temporal resolution plays an important role in speech intellegibillity and in age-related performance decay.

At a resolution of 1 ms, we loose roughly half of the information carried by the spike

trains compared to the finest temporal resolution. This dramatic loss in information rate suggests that spike timing serves as an important mechanism for information transmission. The efficiency of neural coding is fairly high. Depending on the temporal resolution used for the binning process, efficiency ranges from 60% to 90%. Such a high efficiency seems to show that there is important information for speech, which is yet neglected in ASR systems. It is not fully clear to us whether this information is relevant for the discrimination of speakers, to sound localization, or to better speech recognition in noise.

Our calculation provides a reference for technical applications like automatic speech recognition. Present technology relies almost exclusively on short time spectra with coarse temporal resolution (usually 10 ms). As shown in Figure 5.6, we expect that this process destroys at least 85% of the information coded in the human auditory system. It is unlikely that no important phonemic information is destroyed by discarding more than 85% of the temporal information coded by onset neurons. In summary, we suspect that fine-grained temporal information might carry more useful information to recognize speech than anticipated, especially in noisy conditions.

The experiments so far have been performed on a limited set of input stimuli. As a future work, we would like to apply the algorithms to a variety of stimuli such as different vowels, consonants and noises. We also expect to perform a similar analysis on the output of the auditory fibres. This is more challenging, because the variability seen in auditory nerve fiber responses is much larger compared to onset neurons, which work like coincidence detectors on a larger population of ANFs. Comparing the information transmission before and after the processing of Onset Neurons can lead to a better understanding how they code information.

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (reference numbers 01GQ0441 and 01GQ0443).

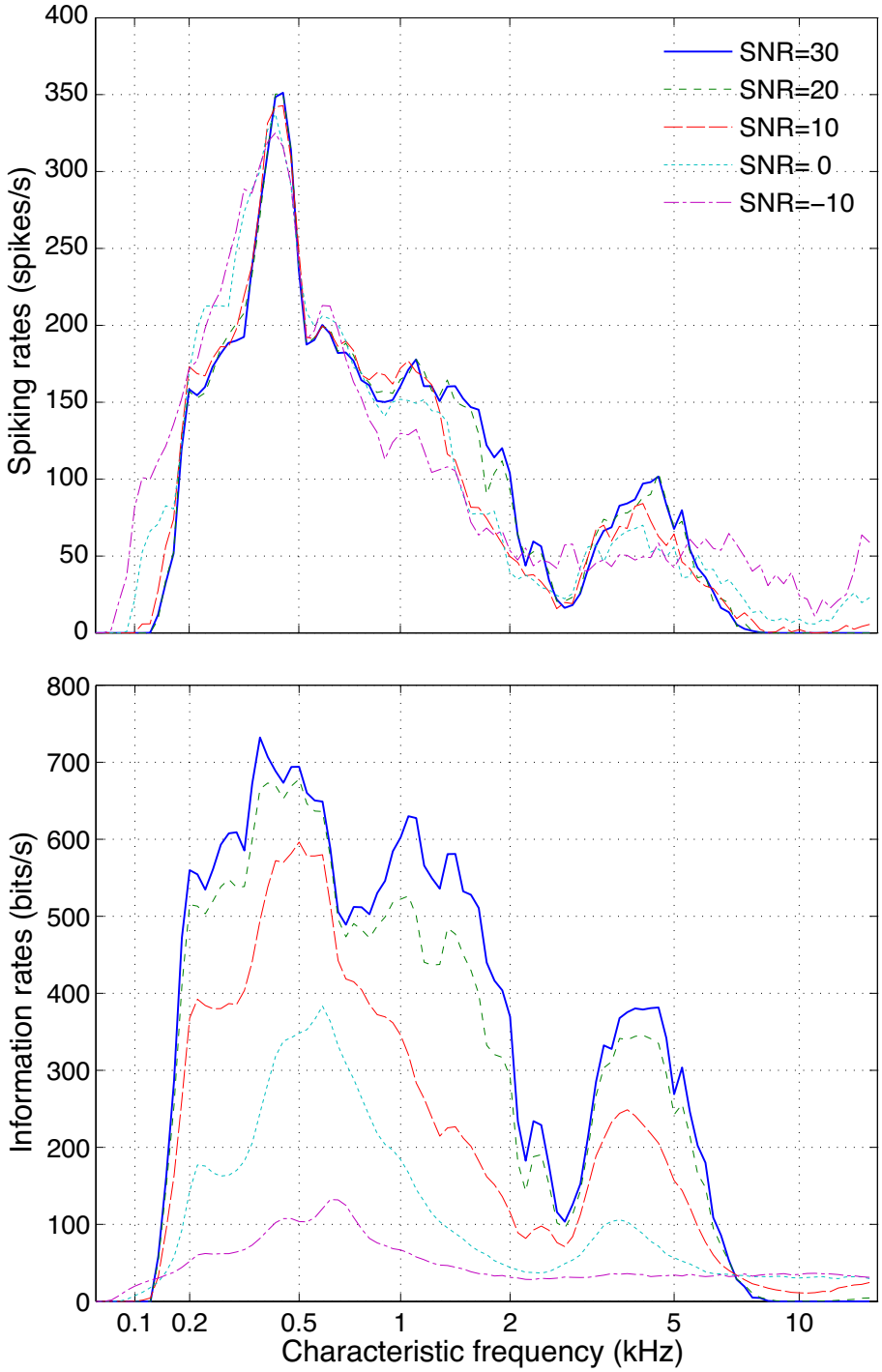

Figure 5.8: Information distributed over frequency channels (lower panel) and average spiking rates (upper panel) for signals with different SNRs (-10 dB, 0 dB, 10 dB, 20 dB, and 30 dB). The input stimulus is vowel /*ei*/, 75 dB(A), added with different levels of pink noise, which is regenerated for every trial. Word duration is 10 ms with 0.25 ms resolution.

6 Analysis of Different Neurons with the Information Theory Approach

Abstract[1)]

The auditory pathway through the ventral cochlear nucleus (VCN) diverges through stellate, bushy and octopus cells to take part in intermediate integrative circuits and transfer information to the inferior colliculus. In this chapter, we focused on analyzing these neurons in VCN, which get direct input from primary auditory nerve fibers (ANF). We generate spike trains of the ANFs and VCN neurons with the inner ear model and calculate the transmitted information using vowels (both real and synthesized) as input stimuli. For ANFs, transmitted information is the highest in the frequency range of $200 - 500\,\mathrm{Hz}$, and decreases towards higher frequencies, due to the degrading temporal precision of the spikes. A single stellate neuron is able to transmit a large portion (up to 66%) of information transmitted by five of its innervating ANFs. Due to their slow membrane time constant the information rate decreases even faster with CF compared to ANFs. The spectral information of the sound signals is well reflected in the rate-place code of ANFs and VCN neurons. Octopus neurons and bushy cells have much faster membrane time constants than stellate neurons. They fire preferably to stimulus onsets or amplitude modulated signals and are very precise in time. Analysis with synthesized vowels showed that, unlike stellate neurons, the firing rates and information rates of octopus neurons and bushy neurons do not match the spectral distribution of the sound signal. They have higher information rates than stellate neurons, even when spiking rates are lower, indicating that they code temporal information rather than a pure spectral energy distribution.

[1)] A modified version of this chapter has been published in Interspeech 2007.[Wang and Hemmert, 2007]

6.1 Introduction

Most of the acoustic information arrives at the brain stem of mammals through auditory nerve fibers that form a single, tonotopically organized pathway. In the synaptic connection of ANFs with different groups of principle cells, the auditory pathway branches into multiple, parallel descending pathways. Pathways through the VCN diverge through bushy, stellate and octopus cells, which then transfer information to higher processing stages of the brain by trains of spikes. How these pathways contribute to the fundamental biological tasks such as sound localization and the coding and processing of speech sounds is only partially understood. Therefore describing and quantifying information carried by these different neurons to higher stages of neural processing is essential for understanding the underlying mechanisms of acoustic processing.

Previous studies have revealed some properties of different neurons. Octopus cells are contacted by many auditory nerve fibers, each providing a very small depolarization (less than 1 mV). Summation of synaptic input from multiple fibers is required for an octopus cell to reach threshold. In firing only when synaptic depolarization exceeds a threshold rate, octopus cells fire selectively when synaptic input is sufficiently large and synchronized. Octopus cells convey features of sound that are critical for the recognition of natural sounds including speech. They code the presence of acoustic transients, periodicity, and direction of frequency sweeps in their temporal firing patterns [Ferragamo and Oertel, 2002b]. The bushy cells that we modeled differ from octopus cells by a hyperpolarization-activated cation current I_h. They have similar properties as that of octopus cells. There is strong evidence that pathways through bushy cells and their targets in the medial and lateral superior olivary nuclei contribute to the localization of sound in the horizontal plane [Oertel, 1999, Joris and Yin, 1998].Stellate cells, in contrast, depend much less on the rate of depolarization. Unlike bushy and octopus cells, stellate cells fire continuously when given a constant stimulus. Although stellate cells fire to the onset of tone stimuli with temporal precision, the timing of subsequent spikes is independent of the phase of the sound. With their supra-threshold inputs, stellate cells are poised to act as relay cells from auditory nerve fibers to higher processing stages.

In this chapter, we utilized the information theory to evaluate the different types of modeled onset neurons. The objective was to understand more about the mechanisms underlying human sound processing from an information transmission perspective.

6.2 Methods

The VCN neurons were modeled as in Chapter 3 and connected to ANFs from the inner ear model (compare Figure 2.1). The parameters of the ion channels of the neurons were taken from Rothman and Manis [Rothman and Manis, 2003a]. All conductances and time-constants were corrected to a body temperature of 38°.

According to Ferragamo et al. [1998], stellate cells in my study are driven by 5 supra-

threshold [Ferragamo and Oertel, 2002b] synaptic inputs from high spontaneous rate auditory nerve fibers (spontaneous rate: 30 spikes/s). We used 60 sub-threshold synaptic inputs for octopus neurons such that synchronous firing of six synaptic inputs generated an action potential [Oertel et al., 2000]. Therefore, synaptic inputs to octopus neurons are sub-threshold and they act as coincidence detectors. There is a plethora of publications reporting on the properties and parameters of these neurons. We used up-to-date empirical data for the modeling from Rothman and Manis [Rothman and Manis, 2003a]. Table 6.1 shows the parameters that we used for modeling the different onset neurons.

	Model Type		
	Stellate	Bushy	Octopus
$\overline{g}_{Na}, nS$	1000	1000	1000
$\overline{g}_{HT}, nS$	150	150	150
$\overline{g}_{LT}, nS$	0	200	600
$\overline{g}_{h}, nS$	0.5	20	40
$\overline{g}_{lk}, nS$	2	2	2
$\overline{g}_{E}, nS$@38^{o}C	33	8.5	24.5
# of synapses	5	60	60
type of synapses	suprathreshold	subthreshold	subthreshold

Table 6.1: Parameters of different neurons.

The information calculation approach was implemented as described in Chapter 5.3. We took the direct method which makes no assumption about the underlying mechanism of information processing. We mainly looked into the information distribution along frequency channels. In order to speed up the calculation, we used a temporal resolution of 1 ms, which was slightly worse than in the previous chapter. Doing so, we lost roughly half of the information. Since the main objective was to compare different neurons instead of quantifying the exact amount of information transmitted, using a coarser resolution did not affect the conclusions, which we confirmed by calculating information at different resolutions for some cases.

We used two different stimuli in the calculation: an /ei/ spoken by a female speaker (ISOLET fcmc0-A1, scaled to 70 dB(A) SPL) and a synthesized vowel /i/ generated with Slaney's auditory toolbox [Slaney, 1998]. The synthesized vowel has a stable spectrum, therefore it is easy to compare the neuronal responses of the auditory model with the power spectrum of the signal.

6.3 Results

In Figure 6.1 we showed the responses of auditory nerve fibers and stellate neurons, which receive input from auditory nerve fibers with different characteristic frequencies.

Input stimulus was the synthesized vowel /i/ (pitch frequency: 220 Hz; F1: 270 Hz; F2: 2290 Hz; F3: 3010 Hz; sound intensity: 70 dB(A)) generated by Slaney's auditory toolbox for Matlab [Slaney, 1998]. As the artificial vowel is (almost perfectly) periodic, its spectrum is comprised of the harmonics of the pitch frequency (indicated by crosses in Figure 6.1b),c),e) and f)). Panel d) plots the raw spike patterns of the stellate neurons.

The spectrum of the speech sound is reflected rather accurately by the spike rates of both auditory nerve fibers (panel b) and stellate cells (panel e). In the region of the pitch frequency and slightly above, ANFs fire with a probability of about 80% per cycle. This precision is improved by stellate cells, which fire one action potential per stimulus cycle with a probability of slightly more than 90%. The property that auditory neurons fire preferentially one spike per stimulus cycle limits their maximum rate in this region.

Individual frequency components of the stimulus are separated in the spike-rate contour of both stellate cells and ANFs, in this example up to the seventh harmonic (1540 Hz). Even higher frequency components merge together and the rate contours follow the envelope of the spectrum, coding the second and – in this example with less salience – third formant. For the second and fourth harmonics the spike rate contour peaked due to the coding of second and fourth harmonics by the auditory nerve fibers and the stellate cells. Note that the peaks of the rate contours are slightly shifted relative to the spectral peaks of the speech sound. This effect appears because the CF axis is defined at threshold levels and excitation patterns are shifted towards higher frequencies at higher sound levels in the mammalian inner ear [Dallos, 1992, Holmberg and Hemmert, 2004]. The spontaneous activity of the high-spontaneous rate ANFs we used here was about 30 spikes/s. The stellate cells are innervated by five high-spontaneous rate auditory nerve fibers. The synapses are supra-threshold, which means that a spike from one auditory nerve alone could elicit an action potential. This caused a high spontaneous activity of the modeled stellate cells of about 80 spikes/s.

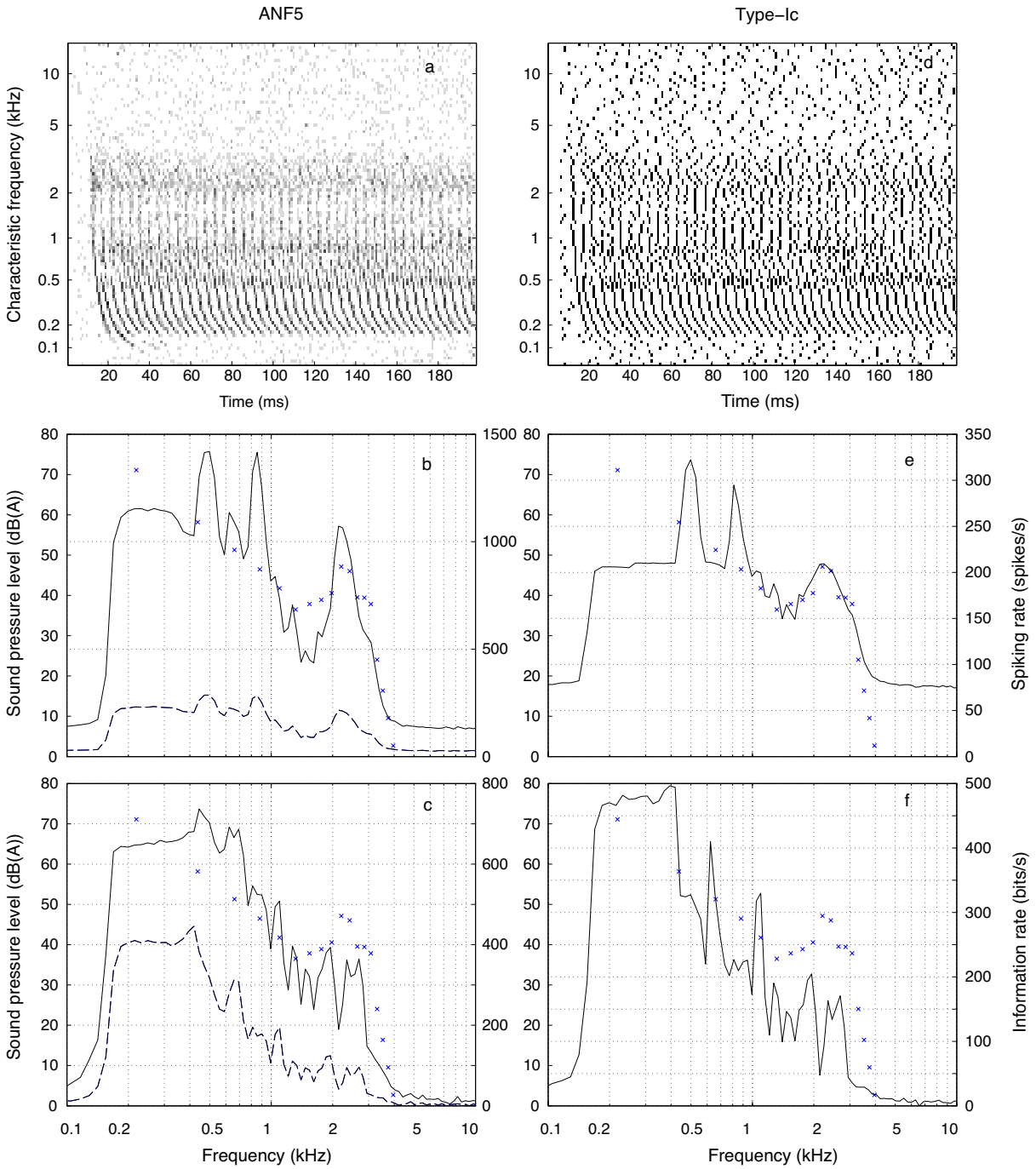

Figure 6.1: Spike trains along the length of the cochlea (first row), averaged spiking rates (second row) and transmitted information rate (third row) from auditory nerve fibers (left column) and stellate neurons (right column). Note that information rates are plotted for a single stellate cell and five ANFs innervating the stellate cell. Information rate and spiking rate of a single ANF (dashed lines) were also plotted for comparison. For reference, rows two and three show the A-filtered levels of the harmonics of the artificial vowel /i/ used for stimulation (pitch frequency: 220 Hz).

The information rate of ANFs and stellate cells is plotted in Figure 6.1c) and f). Note that for comparison we calculated the transmitted information of five ANF spike trains, which excited a single stellate cell. For ANFs, transmitted information was approximately constant (about 650 bits/s) in the range from 200-400 Hz and then decreased rapidly as a function of their characteristic frequency. As for the spike rate, information rate peaked at harmonics of the pitch frequency and at the formant regions of the vowel. For the stellate neurons, information rates were always below the values of five ANFs. Information decreased even more rapidly as a function of their characteristic frequency compared to the values of ANFs. This is due to their long membrane time constant (7 ms [Rothman and Manis, 2003a]), which impairs precise phase-locking at higher frequencies.

We also plotted the spike rate and information rate for one single ANF. With one single ANF, hence one fifth the rate of 5 ANFs, the total amount of information transmitted was 43% compared to that of 5 ANFs altogether, which clearly indicates the redundancy of information in different ANFs. Stellate cells integrate spikes from the 5 ANF and transfer around 54% of the incoming information to higher stages.

Figure 6.2 and 6.3 show spiking rates and information rates of different neurons. Octopus and bushy cells show little or none spontaneous activity as they rarely fire when given spontaneous ANF activity as input. For octopus and bushy cells, it is not necessary that they have high spiking rates in regions where energy is high (e.g. the third formant region). In the first formant region, all neurons have similar spiking rates while bushy and octopus cells have higher information rate than stellate cells since they are more accurate in time. In the frequency range between 0.5 kHz and 1 kHz, octopus cells have the highest rate, which is almost twice as much as the bushy cells. This can be attributed to their very fast membrane time constant which gives them the ability to fire at multiples of the pitch frequency. Stellate cells have higher rates than bushy cells in the frequency range between 500 Hz and 1 kHz. They also have the highest rate at frequencies above 1 kHz. This is mainly due to the fact that the stellate cells code energy rather than the temporal structure, therefore do not suffer from the blurry temporal structure of the spikes from auditory nerve fibers in high frequency regions (see Figure 6.1 a).

The information rate of octopus cells were the largest in all frequency ranges, even when their spiking rates were considerably lower than that of stellate cells, which clearly showed that octopus cells transmit much higher information per spike. Unlike stellate cells, the contour of the information rate for octopus cells and bushy cells didn't follow the energy level of the harmonics. At frequencies below 1.5 kHz, the information rate contour remained fairly flat at a very high level in comparison to ANFs and stellate cells, indicating that they were coding other information besides energy. At frequencies above 1.5 kHz, the information rates of octopus cells and bushy cells decreased dramatically as the temporal structure of incoming ANF spikes are lost. The high information rate of octopus cells makes them very interesting for speech coding.

In Figure 6.4 we compared the transmitted information of three types of different neurons as a function of temporal resolution. We plotted the channel with a CF of 1045 Hz. At this frequency, all the neurons had similar firing rates, but the information rates differed

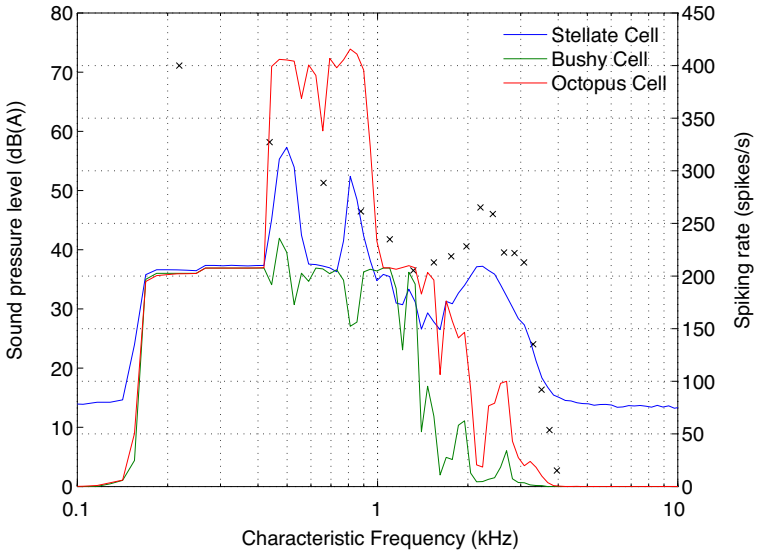

Figure 6.2: Distribution of firing rates of stellate cells, bushy cells and octopus cells. Crosses are A-filtered levels or harmonics calculated using Fast Fourier Transform. It is seen that the firing rates of bushy neurons and octopus neurons do not follow the spectrum, which is nicely revealed in the firing rate of stellate neurons.

considerably (compare Figure 6.2 and Figure 6.3).

The plot shows that the information content of octopus cells and bushy cells are much higher compared to that of stellate cells for this relatively high CF channel, especially for temporal resolutions higher than 1 ms. Note that the octopus cells and bushy cells are innervated by much more ANFs than the stellate cells (60 for octopus and bushy cells compared to 5 for stellate cells). The information rate curves of octopus cells and bushy cells largely overlapped from 10 ms resolution to about 0.4 ms resolution. After that the curves diverged, with octopus cells exhibiing a slightly higher information rate. For the frequency channel that we analyzed, octopus cells and bushy cells had very similar firing pattern and they differed only in temporal scales that were finer than 0.4 ms. The information rate at a temporal resolution of 10 ms, a value which is usually used for automatic speech recognition, was only about 6.9% for octopus neurons and 7.1% for stellate cells compared to the information rate at the highest resolution! This indicates that a huge amount of information present in neuronal spike trains of auditory neurons is discarded when features based on short-term spectra with a temporal resolution of 10 ms are calculated. We therefore conclude that fine-grained temporal information is by an order of magnitude larger than classical spectral features used for ASR. We suspect that this information might be essential for robust speech recognition especially in adverse acoustic conditions.

Figure 6.5 shows the coding efficiencies of the three neuron types. It is not surprising that octopus cells and bushy cells have higher coding efficiencies than stellate cells, as they have higher information rates even when their spiking rates are low. The coding efficiencies

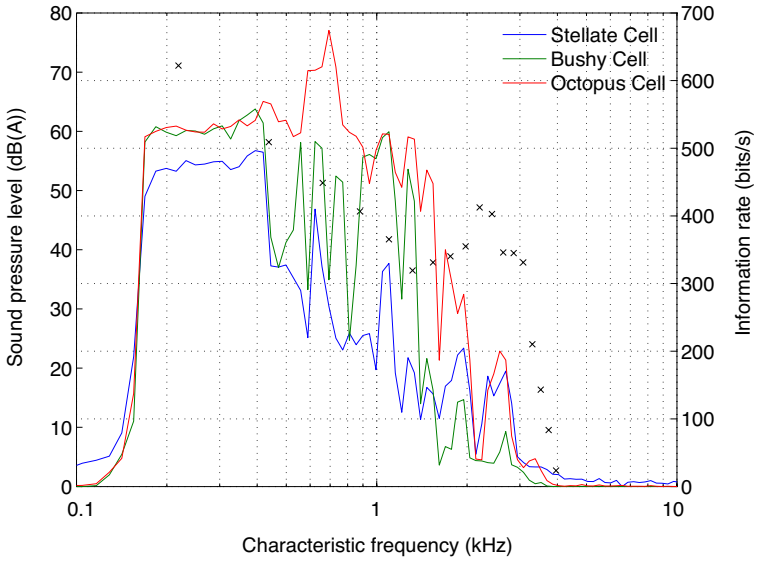

Figure 6.3: Information distribution along frequency channels. Plotted are stellate neurons, bushy neurons and octopus neurons. Stimulus is artificial vowel /i/ at 70 dB(SPL)

of octopus cells and bushy cells were astonishingly high. When resolutions was between 0.4 ms and 10 ms, the efficiencies of coding were fairly high and stable, varying around 95%. The efficiency of bushy cells calculated here is higher than the value that we showed in Chapter. 5.4 as we used synthesized speech signal instead of a real vowel. For a resolution above 0.4 ms, coding efficiency degraded, even though still more information was coded (see Figure 6.4). The efficiency curves of octopus cells and bushy cells showed strong similarities. Octopus cells have slightly higher efficiency than bush cells at most of the temporal resolutions. The coding efficiency of stellate cells is much lower – around 55% from 7 ms to about 0.4 ms and decreases thereafter. The extraordinarily high efficiencies of octopus cells and bushy cells indicate the strength of the temporal code. The ability of reliably locking to signal onsets with very high temporal accuracy allows them to code speech signals efficiently with sparse spikes.

Another interesting comparison of the three neuron types is the average information transmitted by each spike. Figure 6.6 shows the average information per spike at different characteristic frequencies. For stellate neurons, a single spike carries most information at low frequencies. The information-per-spike contour resembles the one of the firing rate curve (plot not shown here). Generally, in regions where energy concentrates, a spike carries in average more information. Bushy cells and octopus cells again show different patterns. In low and high frequency regions, the average information per spike is particularly high. This is related to the firing properties of these two neurons. At low frequencies, bushy and octopus cells phase lock to the carrier frequency while at high frequencies they phase lock to the modulation frequency. In both cases, the spikes have very high temporal precision. At frequencies between 400 Hz and 1 kHz, each spike carries less information. However, due to their extremely fast membrane constant, octopus neurons managed to firing with

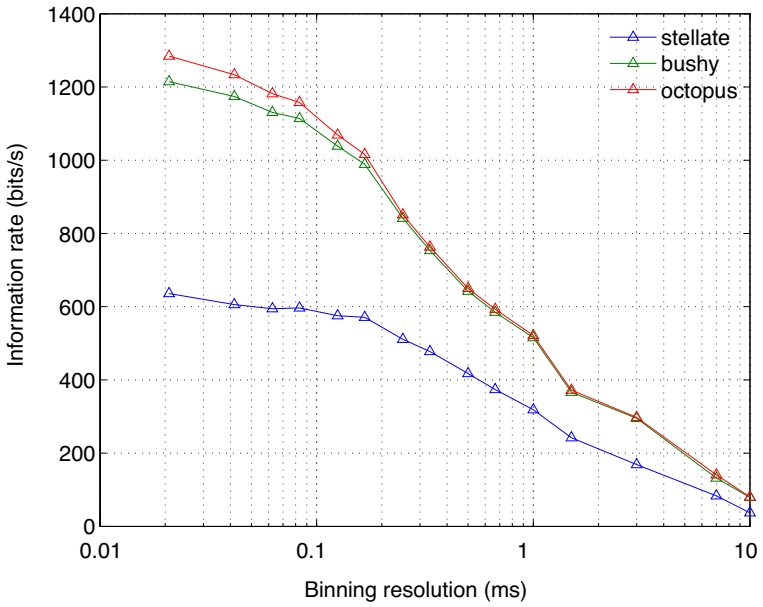

Figure 6.4: Dependence of transmitted information rate on temporal resolution for VCN stellate, bushy and octopus neurons with a characteristic frequency of 1045 Hz.

very high rates, locking to harmonics of the pitch frequency. Therefore, octopus neurons still carry a large amount of information in this region despite carrying only a low average information per spike. Bushy cells on the other side, exhibit a low firing rate, therefore few information.

6.4 Discussion

Although information processing for automatic speech recognition is inspired by auditory sound processing, there are major differences in the way our brain processes sound signals. The most striking disparity is that all information – not only the auditory one – is transmitted and processed using discrete nerve-action potentials. Despite their all-or-none nature, spike trains even from single neurons can code amplitudes using a rate code (compare Figure 6.1b), especially if one considers that usually large groups of neurons are available to code signals. For example, all information about sound signals available to our brain is transmitted by about 32.000 primary auditory nerve fibers. Spectral information is coded by the firing rates of auditory nerve fibers. This well-known rate-place code is based on the mechanical bandpass filtering along the inner ear. The rate-place contour (compare Figure 6.1b) resembles closely the spectrum of the speech signal.

However, the rate-place code is not the only information source available to the brain. A large amount of information is carried by the precise spike timing of the nerve-action potentials. To quantify all the information carried by spike trains of auditory neurons, we applied the framework of information theory (see Chapter 5). Using the so-called direct

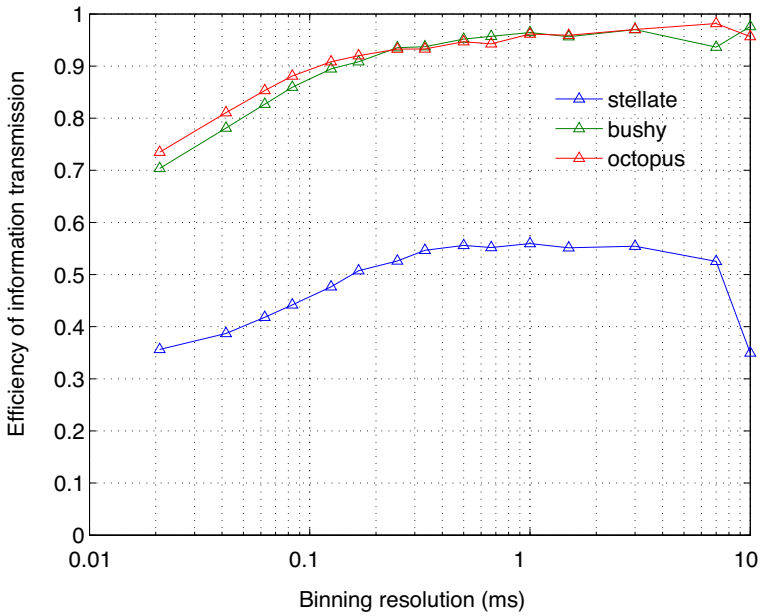

Figure 6.5: Dependence of coding efficiency on temporal resolution for VCN stellate bushy and octopus neurons with a characteristic frequency of 1045 Hz.

method to calculate information in spike trains, we evaluated the transmitted information without making any assumption about the underlying coding strategy. The information rate transmitted by auditory nerve fibers as a function of their characteristic frequency seems to be low-pass filtered compared to the rate-place contour (compare Figure 6.1 b and c). This can at least partly be explained by the phase-locking abilities of these fibers, which decreases for frequencies above about 1 Hz [Rothman and Manis, 2003a]. The decreasing phase-locking ability also entails that the temporal precision of auditory nerve action potentials decreases with increasing CF.

Stellate cells faithfully code spectral information in their rate-place contours, similar to primary auditory nerve fibers (compare Figure 6.1 b and e). With their relatively long membrane time constants (about 7 ms, Rothman and Manis [2003a]) they integrate auditory nerve inputs over time. However, due to this slow membrane time constant their phase locking ability (and as a result the information rate) decreases even faster with CF compared to ANFs (compare Figure 6.1 c and f). Note that the information rate of stellate cells is always lower than the total information rate of the ANFs innervating them, as required by theory. The ratio of information rate transmitted by five ANFs and a stellate cell increases from 1.5 (50 Hz) to about 4 in the range of $2-4$ kHz, reflecting the effects of low-pass filtering due to the stellate cell membrane time constant.

We also investigated information transmission of bushy neurons and octopus neurons, which are also located in the VCN Wang et al. [2006]. Compared to stellate cells, bushy- and octopus neurons even manage to improve temporal precision compared to ANFs Hemmert et al. [2005] by using coincidence detection of at least six of their innervating ANFs. Especially at higher CFs, these neurons are able to transmit much larger information

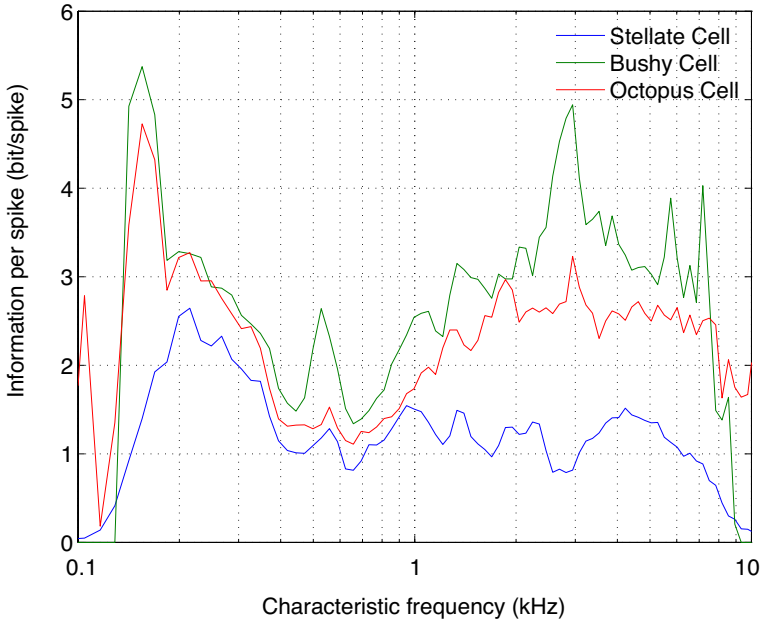

Figure 6.6: Information transmission per spike at different frequency channels for the stellate, the bushy and the octopus neurons.

rates than stellate cells due to their reliable firing, high firing rates of up to 800 spikes/s and the extreme temporal precision of their spikes.

For bushy and octopus neurons, the contour of information rate as a function of frequency channels do not follow the level of harmonics of the input signal. At low frequencies, the information rate remains at a very high level, due to their precise phase locking to the input signal. Bushy and octopus cells almost always have higher information rates than stellate cells, which corroborates that they also carry temporal information besides information about the energy, which is the case for stellate cells. Bushy- and octopus cells exhibit high firing rates at frequencies where phase locking is prevalent (almost irrelevant of energy), thus transforming temporal information into a rate code. At high frequencies (from 2 kHz to 8 kHz), a single spike of bushy- or octopus neurons also carry very high information rates as the precicely phase lock to amplitude modulations present in this frequency range.

We used different temporal resolutions to quantify the information rates of neurons. Usually, bushy and octopus neurons carry more information than stellate cells. They also show a steeper increment when temporal resolution increases. For all types of neurons, using a temporal resolution of 10 ms leads to information loss of more than 90% percent compared to the highest resolution we were able to use in the calculations. Coding efficiencies of bushy neurons and octopus neurons are fairly high and decrease when temporal resolution increases. Stellate neurons' efficiency of coding, in comparison, are considerably lower.

Octopus neurons and bushy neurons are very similar types of neurons. They differ from

each other by a hyper-polarization-activated cation current I_h. Octopus neurons exhibit higher information rates, efficiencies, and firing rates compared to bushy cells. The most obvious difference between these two types of neurons lies in the frequency range between 500 Hz to 1 kHz. In this region, octopus neurons have much higher firing rates than bushy cells (almost twice as high). Octopus neurons also have higher information rates. The efficiencies of coding with different resolutions are approximately the same (calculated for CF of 1045 Hz, using synthesized vowel as input). However, bushy neurons carry more information per spike in average. One possible reason is that bushy neurons have slower membrane constants and do not fire at higher harmonics of the fundamental frequency, therefore their spike might carry less redundant information.

In a nutshell, we applied information theory on different neuron types mainly to see what aspects of information they code. The results clearly show that stellate cells mainly code information about the energy of the spectrum. Bushy- and octopus cells on the other hand, exhibit much higher information rates and coding efficiencies, due to their ability to precisely code temporal information. Temporal information in the speech signal was converted by octopus neurons into a rate code and can be utilized by automatic speech recognizers. In Chapter 4, we also showed how octopus neurons can help improve automatic speech recognition for vowels.

7 Summary and Outlook

The primary goals of this thesis were to understand speech coding in the peripheral human auditory system and to identify principle properties of the information processing by the neural system. We therefore built a framework consisting of a realistic peripheral human auditory model, a speech recognition system for qualitative analysis and an information theory based approach for quantitative assessment. The elements of the framework focus on different properties of the model – and speech coding in general. They complement each other and taken together, they provide a useful tool for continuously improving the auditory modeling for speech recognition system and understanding the underlying principles of auditory coding.

7.1 Overall Discussion and Conclusion

7.1.1 Modeling

The outstanding performance of human auditory system motivates our effort of modeling it and applying it for automatic speech recognition. While traditional ASR systems rely almost exclusively on short term spectral features, the human auditory system exploits spike trains of auditory nerve fibers to code the spectral, temporal and spatial information, which is processed at higher levels in the auditory pathway.

This thesis presents a realistic model of the human auditory system. The system consists of a simplified outer ear and middle ear model, a model of inner ear hydrodynamics followed by a compression stage, the sensory cells and the auditory nerve fibers. Special efforts were made to tune the parameters of the BM model to replicate the human hearing threshold, filter bandwith and compression. After tuning, the model output faithfully replicated latest psychoacoustic measurements of human threshold tuning curves [Shera et al., 2002, Oxenham and Shera, 2003] and the dynamic range compression up to 80 dB [Lopez-Poveda et al., 2003]. The inner ear model was then connected to a model of inner hair cells and auditory nerve fibers [Sumner et al., 2002, Meddis, 1986, 1988]. The realistic tuning threshold and the dynamic compression range were maintained over the whole range of audible frequencies.

We further developed the synaptic model to achieve a more realistic offset adaptation of the auditory nerve response. Offset adaptation completely suppresses spontaneous activity in the primary auditory nerve for a few milliseconds after the stimulus offset, an

effect which is not predicted by conventional pool models of synaptic transmission [Sumner et al., 2002, Meddis, 1988]. We followed the proposal of Zhang and Carney [2005] and added a shift value in the pool model to reproduce physiological measurements of offset adaptation. The enhanced offset adaptation considerably improved phase locking and amplitude modulation coding in the auditory nerve. It was also crucial for the onset neurons in the next processing stage. The onset neurons were able to lock to the amplitude modulation of speech signal at high frequency range, whereas they were blocked when there was no enhanced offset adaptation [Wang et al., 2008].

Extensions of the model also covered neurons in the first processing stage after the cochlea, the auditory brainstem or cochlear nucleus. Cochlear nucleus onset neurons (two types of onset neurons were considered, so-called octopus neurons and bushy cells) and stellate neurons were modeled using single compartment models with Hodgkin-Huxley-type ion channels and conductance-based synapses. Onset neurons fire preferably to the signal onset with very high firing probability, even when the auditory nerve fibers that innervate them are at saturation. The octopus neurons were able to phase lock to the input stimuli up to about 800 Hz. At characteristic frequencies above 1 kHz, onset neurons phase lock on amplitude modulations of the input stimuli. They followed to amplitude modulated signals up to 1 kHz with entrainment. Onset neurons extracted to the temporal structure of speech signals rather than their spectral intensity. Therefore they code temporal information in auditory nerve spike trains and convert them to a rate code. We applied the spike-triggered reverse-correlation technique to analyze the underlying processing capabilities of the modeled onset neurons, and found they perform band-pass filtering on incoming signals and the most sensitive part covered the range of fundamental frequencies of speech and music. The nonlinear properties of these neurons are responsible for the fact that they react preferentially to speech signals rather than sinusoidially amplitude modulated signals.

In summary, our model covers all important processing stages in the inner ear and selected neurons in the auditory brain stem. It therefore can serve as the basis for further studies of speech coding and information processing.

7.1.2 Automatic Speech Recognition

There have been many efforts in studying the encoding of speech in the human auditory system from various aspects. Early works focused on the encoding of steady-state vowels, synthetic consonant-vowel syllables, spoken syllables etc. These studies only investigated well defined speech properties and did not provide a qualitative measure of how well a hypothetical speech encoding scheme really works. This thesis exploits automatic speech recognition as a tool to qualitatively evaluate the principle of the rate-place code. Speech coding in the auditory system was tested with a complex but realistic scenario: speech in noise. Speech recognition tasks were carried out in many different modes, e.g. with and without enhanced offset adaptation and with two different testbeds: HTK using Gaussian Mixture Models for acoustic modeling and an MLP testbed using multi-layer perceptrons

for the acoustic modeling.

Speech recognition using the auditory model achieved comparable performance as the speech recognition baseline, which used MFCC ASR features for the recognition. Best recognition performance was achieved using the auditory model with the enhanced offset adaptation and the MLP testbed. The enhanced offset adaptation not only generated realistic auditory nerve outputs per se, but also improved speech recognition performance. When tested on the noisy alphabet using features derived from auditory nerve fibers, the enhanced offset adaptation improved average speech recognition rates for the HTK and MLP testbeds by 12.5% and 19.6%, respectively. The enhanced offset adaptation was also crucial for the octopus neurons to spike properly. It improved vowel recognition using features derived from the octopus neurons by 37.1% for the HTK testbed and 38.4% for the MLP testbed.

The acoustic modeling using multi-layer perceptrons makes less assumptions about speech compared to acoustic modeling using GMMs. In most of the cases, MLP testbed substantially outperformed the HTK testbed, except when MFCCs were used for the vowel subset. This is consistent with past results on MSG auditory features. MLPs also make combining features from different auditory processing stages much easier. The multi-stream approach using combined features from ANFs and octopus neurons outperformed each single feature in vowel recognition tests.

We also studied the discrimination ability and robustness of the auditory system over a large dynamic range. Previously, it was shown that a rate-place representation is not robust over the range of levels encoded by humans, due to nonlinear effects such as discharge rate saturation and suppression [Young and Sachs, 1979]. More recent studies have suggested that the rate-place code of vowels might be sufficient to explain the dynamic range of the human vowel perception [Conley and Keilson, 1995]. Our study also showed that the rate-place coding is robust against the increasing level of the speech. The recognition performance for multi-conditional training and test setup degraded only moderately. Recognition result dropped only by 10.7% in total when SPL increased from 45 dB to 105 dB. The offset adaptation plays an very important role for the robustness against speech level. Without the enhanced offset adaptation, the corresponding recognition performance dropped at a dramatic rate of 9.4% per 10 dB. The average recognition rate dropped by 56.4% when the speech level increased from 45 dB to 105 dB. The robustness of the result using auditory features resembles human performance (see Figure 4.7 and Figure 4.8), even though there is still a big gap in terms of absolute recognition performance. The reason for this gap might be that speech recognition features discard fine temporal information that is coded by the neural spike trains.

7.1.3 Analysis of Auditory Spike Trains with Information Theory

Automatic speech recognition provides a tool to evaluate how well speech can be discriminated using selected features. It shows that the rate-place code of auditory nerve fibers and octopus neurons provide an effective representation of speech signals. However, it is

very hard to draw any direct conclusions about the neural code based on ASR results. With ASR, we could only investigate the rate-code, while the accuracy of individual spike timing represent a second, if not the most important mechanism of information transmission along the neural pathway [Rieke et al., 1997]. Therefore we developed an approach based on information theory to analyze the neural code.

We used the direct method for estimating the information coded in neuronal spike trains. The direct method does not make any assumptions on the neural code and therefore preserves all the information present in the neural spike trains. We found out that information rate is not constant for increasing word lengths, which was in contrast to earlier studies. Instead, the transmitted information saturates exponentially, with 90% of the information transmitted within the first 45 ms (vowel /ei/ from a female speaker in the ISOLET database). For louder signals the information was transmitted faster.

Results showed that onset neurons code information with a very high rate. We found the information rate could be as high as 1050 bits/s, which corresponds to an information content of about 5.8 bits/spike. Our results showed that temporal information is very important for neural coding. The temporal resolution of onset neurons is very high from an information transmission point of view. ONs code temporal information up to the highest resolution we could investigate: 20 μs. This high resolution is required for sound localization in the horizontal plane, where ONs are believed to play a major role. The efficiency of neural coding is fairly high. Depending on the temporal resolution used for the binning process, the efficiency ranges from 60% to 90%. The high resolution and efficiency seem to indicate that for speech a large portion of information is coded with high temporal precision.

With a temporal resolution of 0.25 ms, about 80% of the information is transmitted compared to the information at the highest resolution. However, even the (20%) information coded with higher temporal precision seems to be important. It was found that the loss of these information plays an important role in age-related performance decay in speech intelligibility. At a resolution of 1 ms, roughly half of the information carried by the spike trains was transmitted. This dramatic loss in information rate suggests that spike timing serves as an important mechanism for the information transmission. However, it is not very clear whether the information coded with fine resolution is relevant for the discrimination of speakers, for sound localization, for better speech recognition in noise or for all of theses tasks.

The automatic speech recognition system that we implemented used a temporal resolution of 10 ms, which is also common for many other speech recognition systems. Our analysis using the information theory based approach showed that with this temporal resolution, more than 85% percent of the information coded by the human auditory system was destroyed at this temporal resolution. It is very unlikely that the lost information does not contain important phonetic information. We suspect the loss of fine-grained information from the spike trains might contribute to the performance gap in speech recognition between the modeled auditory system and the real human auditory system.

We also showed how information was distributed over the frequency channels in the au-

ditory system and how robust the coding was in noise. For a voiced speech signal with added pink noise, the information content decreased significantly at 0 dB SNR: Compared to clean speech signal, the decrements were 85% in the 1.5 kHz channel and 73% in the 260 Hz channel. Information concentrates at formant regions of the input speech. The distribution pattern of the information is well preserved at high SNRs (10 dB to 30 dB) and degrades considerably at an SNR of 0 dB.

We exploited the approach based on information theory to analyze different types of neurons. We compared the spiking rates and information rates of auditory nerve fibers and stellate neurons with the spectrum of the input speech signal. Stellate neurons faithfully code spectral information in the rate-place contour, similar to the primary auditory nerve fibers. The spectrum of the speech signal is reflected rather accurately by the spike rates of both auditory nerve fibers and stellate neurons. The individual frequency components of the stimulus were separated on the spiking rate contour up to the seventh harmonic (1540 Hz). Even higher frequency components merge together and the rate contours follow the envelope of the sprectrum.

Information rates also peak at harmonics of the pitch frequency and at the formant region of the vowel. The information transmitted by ANFs as a function of their characteristic frequency seems to be low-pass filtered compared to their rate contour. This is at least partly due to the decreasing phase-locking capability of these fibers at frequencies higher than 1 kHz. Stellate neurons have long membrane constants of about 7 ms. They integrate auditory nerve inputs over time. The information rate of stellate neurons is always lower than the total information rate of the ANFs innervating them, which is required by the data processing inequity. The ratio of information rate transmitted by five ANFs compared to a stellate neuron increases from 1.5 at 50 Hz to about 4 in the range of 2–4 kHz, which reflects the low-pass filtering of the stellate neurons.

The contours of both spiking rate and information rate from bushy and octopus neurons do not follow the level of harmonics of the input signal. In the frequency range below 800 Hz, onset neurons lock on the stimulation frequency. For voiced speech signals, therefore their firing rate increases almost stepwise with every higher harmonic component of the fundamental frequency. At frequencies above about 1 kHz, they lock on the amplitude modulation of the signal with high temporal precision the information rates remain at a very high level. Bushy and octopus neurons almost always code higher information rates than stellate neurons. They also exhibit higher a information per spike. Our results indicate that bushy and octopus neurons code temporal information besides the information about signal energy, which is mainly the case for the rate-place code of stellate neurons. Bushy and octopus neurons on the other hand, were capable of coding precise temporal information. They carry very high information rates and exhibit high coding efficiency.

7.2 Future Work

In this thesis we presented a detailed auditory model which generates very realistic outputs. However it can certainly be improved. One possible improvement of the modeled auditory system would be the inclusion of the efferent system. This feedback mechanism seems to control both the mechanical vibration of the basilar membrane and the encoding of these vibrations into action potentials in the auditory nerve. They might be critical in correctly understanding speech encoding in the auditory nerve, in particular when there is background noise.

Automatic speech recognition provides a quantitative tool to evaluate the modeled auditory system. However, we mainly investigated the rate-place code and temporal information provided by onset neurons only indirectly. Onset neurons are believed to extract precise temporal information from the input stimuli and turn the temporal information into a rate-code. In the future it will be interesting to test other coding strategies which directly exploit temporal information in the spike trains. The modeling power of the MLP testbed can also be exploited in future investigations. For example, with the MLP testbed, features extracted from different types of neurons, as well as features using different coding strategy can be easily combined. Such experiments will provide us with more insight in designing robust speech recognition systems which might eventually help reducing the gap between the ASR and the human speech intelligibility.

There is also a lot of potential work to do regarding the information processing of the neural system. The experiments so far have been performed on a limited set of input stimuli. In the future, we would like to apply the algorithm to a variety of stimuli such as different vowels, consonants and noises. Current information calculation is an average value over the whole input stimulus. It would be interesting to calculate how the information is distributed along time. The temporal distribution of information may help us to optimize the feature extraction process for speech recognition. In previous experiments the output spike trains were averaged using Hanning window with fixed window width and fixed advancing step. The information distribution over time could help to optimized the window width and/or the position of the Hanning window. In our study we point out that onset neurons use very high temporal resolution to code information. Using a rawer resolution leads to information loss. However, it is unclear what kind of information is affected first by reducing the temporal resolution. We are very interested in designing experiments for testing the characteristics of the lost information.

Bibliography

E. D. Adrian. *The basis of sensation: the action of the sense organs.* London: Christophers; New York: Norton, 1928. 76

ANSI. Methods for the calculation of the speech intelligibility index. ANSI S3.5-1997, American National Standard Institute, New York, 1997. 71, 72

J. Atick. *Could information theory provide an ecological theory of sensory processing?* Network: Computation in Neural Systems, 3, 213-251, 1992. 7

F. Attneave. Some information aspects of visual perception. *Psychological review*, 61: 183–193, 1954. 7, 78

R. Bal and D. Oertel. Hyperpolarization-activated, mixed-cation current (i_h) in octopus cells of the mammalian cochlear nucleus. 84:806–817, 2000. 37, 39, 43

R. Bal and D. Oertel. Potassium currents in octopus cells of the mammalian cochlear nucleus. 86:2299–2311, 2001. 37, 44

H. Barlow. *The coding of sensory messages. Current problems in animal behavior p.330-360.* Cambridge University Press, 1959a. 7

H. Barlow. *Sensory mechanisms, the reduction of redundancy, and intelligence In Mechanisation of thought processes.* London: Her Majesty's stationary office, p.535-539, 1959b. 7

H. Barlow. *Redundancy reduction revisited.* Network: Computation in Neural Systems, 12, 241-253, 2001. 7

R. C. Beattie and M. J. M. Raffin. Reliability of threshold, slope, and PB max for monosyllabic words. *J. Speech Hear. Disord.*, 50:166–178, 1985. 72

C. Becchetti and L. P. Ricotti. *Speech Recognition Theory and C++ Implementation.* John Wiley & Sons, 2002. 1, 5

S. Becker. *Mutual information maximization: Models of cortical self organization.* Network: Computation in Neural Systems,7,7-31, 1996. 7

G. Békésy. *Experiments in hearing.* McGraw-Hill Book Company, 1960. 5

J. Blauert. *Räumliches Hören.* Hirzel Verlag Stuttgart, 1974. 4

J. Blauert. *Spatial hearing: the psychophysics of human sound localization.* MIT Press, Cambridge, MA, revised edition, 1997. 4, 9

A. Borst. Noise, not stimulus entropy, determines neural information rate. *Journal of Computational Neuroscience*, 14(1):23–31, January 2003. 87

A. Borst and J. Haag. Effects of mean firing on neural information rate. *Journal of Computational Neuroscience*, 10(2):213–221, March 2001. 89

A. Borst and F. Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2:947–957, 1999. 81, 87

A. S. Bregman. *Auditory scene analysis.* MIT Press, Cambridge, USA, 1990. 38

M. Brown and J. Ledwith. Projections of thin (type-ii) and thick (type-i) auditory-nerve fibers in the cochlear nucleus of the mouse. *Hear. Res.*, 1990. 38

L. H. Carney. A model for the responses of low-frequency auditory-nerve fibers in cat. 93:401–417, 1993. 13

L. H. Carney and C. D. Geisler. A temporal analysis of auditory-nerve fiber responses to spoken stop consonant-vowel syllables. 79(6):1896–1914, 1986. 53, 72

G. Chechik. *An information theoretic approach to the study of auditory coding.* PH.D Thesis, Hebrew University, 2003. 8

R. Cole, Y. Muthusamy, and M. Fanty. The ISOLET spoken letter database. Technical Report CS/E 90-004, Oregon Graduate Institute, 1990. URL citeseer.ist.psu.edu/cole94isolet.html. 60, 77

R. Comerford, J. Makhoul, and R. Schwartz. The voice of the computer is heard in the land (and it listens too!)[speech recognition]. *IEEE Spectrum*, 1997. 6

R. A. Conley and S. E. Keilson. Rate representation and discriminability of second formant frequencies for /ε/-like steady-state vowels in cat auditory nerve. 98:3223–3234, 1995. 6, 53, 72, 108

M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Commun.*, 35:141–177, 2001. 38

N. P. Cooper and W. S. Rhode. Basilar membrane mechanics in the hook region of cat and guinea-pig cochleae: sharp tuning and nonlinearity in the absence of baseline position shifts. 63:163–190, 1992. 20

N. P. Cooper and W. S. Rhode. Mechanical responses to two-tone distortion products in the apical and basal turns of the mammalian cochlea. *J Neurophysiol*, 78(1):261–70, Jul 1997. 19

T. Cover and J. Thomas. *The elements of information theory.* New York: Plenum Press., 1991a. 6

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991b. 8, 80

P. Dallos. Low-frequency auditory characteristics: Species dependence. *J. Acoust. Soc. Am.*, 48:489–499, 1970. 11

P. Dallos. Organ of corti kinematics. 4:416–421, 2003. 15

P. Dallos. The active cochlea. *Journal of Neuroscience*, 12:4575–4585, Dec 1992. 97

N. Daly. Recognition of words from their spellings: Integration of multiple knowledge sources. Master's thesis, Massachusetts Institute of Technology, 1987. 72

E. de Boer. Auditory physics. Physical principles in hearing theory. I. *Phys. Rep.*, 62: 87–174, 1980. 10

E. de Boer. Auditory physics. Physical principles in hearing theory. II. *Phys. Rep.*, 105: 141–226, 1984. 10

E. de Boer. *The Cochlea*, chapter Mechanics of the cochlea: Modeling efforts, pages 258–317. Springer, New York, 1996. 13

E. de Boer and H. R. de Jongh. On cochlear encoding: Potentialities and limitations of the reverse correlation technique. 63:115–135, 1978. 50

B. Delgutte and N. Y. S. Kiang. Speech coding in the auditory nerve: I. Vowel-like sounds. *J. Acoust. Soc. Am.*, 75:866–878, 1984a. 2, 53, 72

B. Delgutte and N. Y. S. Kiang. Speech coding in the auditory nerve: V. Vowels in background noise. 75, 1984b. 53

L. Deng and C. D. Geisler. A composite auditory model for processing speech sounds. 82 (6):2001–2012, 1987a. 13

L. Deng and C. D. Geisler. Responses of auditory-nerve fibers to nasal consonant-vowel syllables. 82(6):1977–1988, 1987b. 53, 72

G. Ehret. Quantitative analysis of nerve fibre densities in the cochlea of the house mouse (*Mus musculus*). 183:73–88, 1979. 38

D. Ellis. `http://www.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html`, a. 60, 61

D. Ellis. `http://www.icsi.berkeley.edu/~dpwe/respite/multistream/aurora1999.html`, b. 61, 73

D. Ellis. Improved recognition by combining different features and different systems, 2000. URL `citeseer.ist.psu.edu/ellis00improved.html`. 63

M. J. Ferragamo and D. Oertel. Octopus cells of the mammalian ventral cochlear nucleus sense the rate of depolarization. 87(5):2262–70, 2002a. 51

M. J. Ferragamo and D. Oertel. Octopus cells of the mammalian ventral cochlear nucleus sense the rate of depolarization. *J. Neurophysiology*, 87:2262–2270, 2002b. 95, 96

M. J. Ferragamo, N. Golding, and D. Oertel. Synaptic inputs to stellate cells in the ventral cochlear nucleus. *J. Neurophysiology*, 79:51–63, 1998. 95

H. Fletcher and R. H. Galt. The perception of speech and its relation to telephony. 22: 89–151, 1950. 71

C. D. Geisler, L. Deng, and S. R. Greenberg. Thresholds for primary auditory fibers using statistically defined criteria. 77(3):1102–1109, 1985. 21

O. Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. 2:115–132, 1994. 54

B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. 47(1–2):103–138, 1990. 16

D. A. Godfrey, N. Y. S. Kiang, and B. E. Norris. Single unit activity in the posteroventral cochlear nucleus of the cat. 162:247–268, 1975. 38

J. M. Goldberg and P. B. Brown. Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. *J Neurophysiology*, 32:613–636, 1969. 28

N. L. Golding, D. Robertson, and D. Oertel. Recordings from slices indicate that octopus cells of the cochlear nucleus coincident firing of auditory nerve fibers with temporal precision. 15:3138–3153, 1995. 38

N. L. Golding, M. Ferregamo, and D. Oertel. Role of intrinsic conductances underlying transient responses of octopus cells of the cochlear nucleus. 19:2897–2905, 1999. 37

J. L. Goldstein. Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering. 49:39–60, 1990. 54

D. D. Greenwood. A cochlear frequency-position function for several species – 29 years later. *J. Acoust. Soc. Am.*, 87:2592–2605, 1990. 16

D. Harris and P. Dallos. Forward masking of auditory nerve fiber responses. *J. Neurophysiol*, 42(4):1083–1107, 1979. 25

M. G. Heinz, H. S. Colburn, and L. H. Carney. Quantifying the implications of nonlinear cochlear tuning for auditory-filter estimates. 111(2):996–1011, 2002. 16

W. Hemmert and M. Holmberg. A phenomenological model of human auditory processing. *submitted to the Journal of the Acoustical Society of America*, 2009. 9, 47

W. Hemmert, M. Holmberg, and U. Ramacher. Temporal sound processing by cochlea nucleus octopus neurons. *Proc. ICANN 2005, LNCS 3696*, pages 583–588, 2005. 30, 50, 77, 91, 103

M. Holmberg. *Speech encoding in the human auditory periphery: Modeling and quantitative assessment by means of automatic speech recognition.* PhD thesis, Technical University Darmstadt, Darmstadt, Germany, 2007. 9, 47, 76

M. Holmberg and W. Hemmert. An auditory model for coding speech into nerve-action potentials. In *Proc. Joint Congress CFA/DAGA'04*, pages 773–4, Strasbourg, France, 2004. 59, 76, 77, 97

M. Holmberg, D. Gelbart, and W. Hemmert. Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition. *Speech Communication*, 2007. doi: 10.1016/j.specom.2007.05.009. 51, 60, 76

W. J. Holmes and M. J. Russell. Probabilistic-trajectory segmental HMMs. *Computer Speech and Language*, 13(1):3–37, 1999. 55

D. Johnson. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J Acoust Soc Am*, 68:1115–1122, 1980. 22, 29, 30

P. X. Joris and T. C. Yin. Responses to amplitude-modulated tones in the auditory nerve of the cat. *J. Acoust. Soc. Am.*, 1992. 30

P. X. Joris and T. C. Yin. Coincidence detection in the auditory system: 50 years after jeffress. *Neuron*, 1998. 95

N. Y. S. Kiang, T. Watanabe, E. C. Thomas, and L. F. Clark. Discharge patterns of single fibers in the cat's auditory nerve. Technical report, MIT University Press, Cambridge, MA, 1965. 28, 29

R. Klinke and S. Silbernagl. *Lehrbuch der Physiologie.* Georg Thieme Verlag, Stuttgart, 1994. 39

C. Koch. *Biophysics of Computation - Information Processing in Single Neurons.* Oxford University Press, 1999. 40

J. Lazzaro and J. Wawrzynek. Speech recognition experiments with silicon auditory models. *Analog Integrated Circuits and Signal Processing*, 13:37–51, 1997. 54

M. C. Liberman. Auditory-nerve response from cats raised in a low-noise chamber. 63 (2):442–455, 1978. 21

R. Linsker. *Self organization in a perceptual network.* Computer, 21, 105-117, 1988. 7

R. Linsker. *An applicatioin of the principle of maximum information preservation to linear systems.* Advances in neural information processing systems Vol.1,p.186-194, 1989. 7

E. A. Lopez-Poveda, C. J. Plack, and R. Meddis. Cochlear nonlinearity between 500 and 8000 Hz in listeners with normal hearing. *J. Acoust. Soc. Am.*, 113:951–960, 2003. 17, 59, 77, 106

R. Lorente de No. Iii.-general plan of structure of the primary cochlear nuclei. *Laryngoscope*, 1933. 38

R. Manis. ModelDB: CN bushy, stellate neurons. Online, 2004. URL http://senselab.med.yale.edu/senselab/modeldb/ShowModel.asp?model=37857. 43

R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *J. Acoust. Soc. Am*, 79(3):702–711, 1986. 26, 106

R. Meddis. Simulation of auditory-neural transduction: Further studies. *J. Acoust. Soc. Am*, 83(3):1056–1063, 1988. 26, 106, 107

R. Meddis, L. P. O'Mard, and E. A. Lopez-Poveda. A computational algorithm for computing nonlinear auditory frequency selectivity. *J. Acoust. Soc. Am.*, 109:2852–2861, 2001. 13

S. Mehrgardt and V. Mellert. Transformation characteristics of the external human ear. 61:1567–1576, 1977. 21

M. I. Miller and M. B. Sachs. Representation of stop consonants in the discharge patterns of auditory-nerve fibers. 74:502–517, 1983. 53, 72

N. Morgan and H. Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12:24–42, 1995. 2, 55, 57, 61

H. Müsch and S. Buus. Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance. 109:2910–2920, 2001. 71

S. S. Narayan, A. N. Temchin, A. Recio, and M. A. Ruggero. Frequency tuning of basilar membrane and auditory nerve fibers in the same cochleae. *Science*, 282:1882–1884, 1998. 20

I. Nelken, G. Chechik, T. D. Mrsic-Flogel, A. J. King, and J. W. H. Schnupp. Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. *Journal of Computational Neuroscience*, 19(2):199–221, October 2005. 80

I. Nemenman, F. Shafee, and W. Bialek. *Entropy and inference, revisited.* Advances in neural information processing system 14. Cambridge, MIT Press, 2002. 8

I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69:056111, 2004. URL doi:10.1103/PhysRevE.69.056111. 82

D. Oertel. The role of timing in the brain stem auditory nuclei of vertebrates. *Annual review of physiology*, 61:497–519, 1999. 95

D. Oertel, R. Bal, S. Gardner, P. Smith, and P. Joris. Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. *PNAS*, 97:11773–11779, 2000. 38, 44, 96

R. Oettinger and H. Hauser. Ein elektrischer Kettenleiter zur Untersuchung der mechanischen Schwingungsvorgänge im Innenohr. *Acustica*, 11:161–177, 1961. 10

G. S. Ohm. Über die Definition des Tones nebst daran geknüpfter Theorie der Sirene and ähnlicher tonbildender Vorichtungen. *Ann. Phys. Chem.*, 59:513–565, 1843. 53

A. J. Oxenham and S. P. Bacon. *Compression: From Cochlea to Cochlear Implants*, chapter Psychophysical manifestations of compression: normal-hearing listeners, pages 62–106. Springer, New York, 2004. 17

A. J. Oxenham and C. A. Shera. Estimates of human cochlear tuning at low levels using forward and simultaneous masking. *J Assoc Res Otolaryngol*, 4(4):541–54, December 2003. 16, 106

L. C. Peterson and B. P. Bogert. A dynamical theory of the cochlea. 22:369–381, 1950. 10

M. K. Pichora-Fuller, B. A. Schneider, E. MacDonald, H. E. Pass, and S. Brown. Temporal jitter disrupts speech intelligibility: A simulation of auditory aging. *Hearing Research*, 223:114–121, 2007. 91

C. J. Plack and V. Drga. Psychophysical evidence for auditory compression at low characteristic frequencies. 113(3):1574–1586, 2003. 17, 20

C. J. Plack and A. J. Oxenham. Basilar-membrane nonlinearity and the growth of forward masking. 103:1598–1608, 1998. 20

I. Pollack and J. M. Pickett. Masking of speech by noise at high sound levels. 30:127–130, 1958. 72

A. Rees and A. Palmer. Neuronal responses to amplitude-modulated and pure-tone stimuli in the guinea pig inferior colliculus, and their modification by broadband noise. *J. Acoust. Soc. Am.*, 1989. 31

W. S. Rhode. Temporal coding of 200% amplitude modulated signals in the ventral cochlear nucleus of cat. 77:43–68, 1994. 38, 47

W. S. Rhode. Neural encoding of single-formant stimuli in the ventral cochlear nucleus of the chinchilla. 117:39–56, 1998. 38

W. S. Rhode and S. Greenberg. *The mammalian auditory pathway: Neurophysiology*, chapter Physiology of the cochlear nuclei, pages 94–152. Springer-Verlag, New York, 1992. 62

W. S. Rhode and P. H. Smith. Encoding of timing and intensity in the ventral cochlear nucleus of the cat. 56:261–286, 1986. 38, 46, 47

F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes, Exploring the Neural Code*. MIT Press, Cambridge, MA, 1997. 76, 81, 109

A. Robert and J. L. Eriksson. A composite model of the auditory periphery for simulating responses to complex sounds. *J. Acoust. Soc. Am.*, 106:1852–1864, 1999. 13

L. Robles and M. A. Ruggero. Mechanics of the mammalian cochlea. *Physiol. Rev.*, 81: 1305–52, July 2001. 11, 17, 19, 20

J. S. Rothman and P. B. Manis. The roles potassium currents play in regulating the electrical activity of ventral cochlear nucleus neurons. *J. Neurophysiology*, 89:3097–3113, 2003a. 23, 24, 30, 37, 39, 41, 43, 77, 95, 96, 99, 103

J. S. Rothman and P. B. Manis. Differential expression of three distinct potassium currents in the ventral cochlear nucleus. 89:3070–3082, 2003b. 39

J. S. Rothman and P. B. Manis. Kinetic analyses of three distinct potassium conductances in ventral cochlear nucleus neurons. 89:3083–3096, 2003c. 39

M. A. Ruggero and A. N. Temchin. The roles of the external, middle, and inner ears in determining the bandwidth of hearing. 99(20):13206–13210, 2002. 9, 16

M. A. Ruggero and A. N. Temchin. Middle-ear transmission in humans: wide-band, not frequency-tuned? 4(2):53–58, 2003. 9

M. A. Ruggero and A. N. Temchin. Unexceptional sharpness of frequency tuning in the human cochlea. 102(51):18614–18619, 2005. 16

M. A. Ruggero, N. C. Rich, A. Recio, S. S. Narayan, and L. Robles. Basilar-membrane responses to tones at the base of the chinchilla cochlea. *J Acoust Soc Am*, 101(4): 2151–63, Apr 1997. 19, 20

M. B. Sachs and E. D. Young. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. 66(2), 1979. 2, 53, 65, 72

M. B. Sachs, H. F. Voigt, and E. D. Young. Auditory nerve representation of vowels in background noise. 50(1):27–45, 1983. 53

S. Seneff. A computational model for the peripheral auditory system: Application to speech recognition research. In *IEEE Proc. ICASSP*, pages 1983–1986, 1986. 54

S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *J. Phonetics*, 16:55–76, 1988. 54

C. E. Shannon. A mathematical theory of communication. *The Bell systems technical journal*, 27:379–423, 623–656, 1948. 78

S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky. Feature extraction using non-linear transformation for robust speech recognition on the Aurora database. In *ICASSP*, 2000. 73

E. A. G. Shaw. Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. 56:1848–1861, 1974. 21

H. Sheikhzadeh and L. Deng. Speech analysis and recognition using interval statistics generated from a composite auditory model. 6:90–94, 1998. 54

C. A. Shera, J. J.J. Guinan, and A. Oxenham. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proc Natl Acad Sci USA*, 99(5): 3318–3323, 2002. 16, 54, 59, 77, 106

S. M. Silkes and C. D. Geisler. Responses of "low-spontaneous-rate" fibers to speech syllables presented in noise. I: General characteristics. 90:3122–3139, 1991. 2, 53

D. G. Sinex and C. D. Geisler. Responses of auditory-nerve fibers to consonant-vowel syllables. 73:602–615, 1983. 53, 72

W. Singer and C. Gray. *Visual feature integration and the temporal correlation hypothesis.* Annual Reviews in Neuroscience, 18, 555-586., 1995. 7

M. Slaney. Auditory toolbox. Technical Report 1998-010, Interval Research Corperation, malcolm@interval.com, 1998. 96, 97

H. Steeneken and F. Geurtsen. Description of the RSG-10 noise database. Technical report, TNO Institute for Perception, The Netherlands, 1988. 60

D. Strelioff and A. Flock. Stiffness of sensory-cell hair bundles in the isolated guinea pig cochlea. 15:19–28, 1984. 18

S. Strong, R. de Ruyter van Steveninck, W. Bialek, and R. Koberle. Entropy and information in neural spike trains. *Phys Rev Lett*, 80:197–200, 1998. 79, 84, 87

H. Strube. A computionally efficient basilar-membrane model. *Acustica*, 58:207–214, 1985. 10, 11, 13

G. A. Studebaker, R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney. Monosyllabic word recognition at higher-than-normal speech and noise levels. 105:2431–2444, 1999. 72

C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis. A revised model of the inner-hair cell and auditory-nerve complex. *J. Acoust. Soc. Am.*, 111:2178–88, May 2002. 15, 16, 18, 26, 29, 59, 106, 107

G. Svirskis, V. Kotak, D. Sanes, and J. Rinzel. Sodium along with low-threshold potassium currents enhance coincidence detection of subthreshold noisy signals in mso neurons. *J. Neurophysiology*, 91:2465–2473, 2004. 37, 49, 91

E. Terhardt. Calculating virtual pitch. 1(2):155–82, 1979. 16, 21

F. Theunissen and J. Miller. Temporal encoding in nervous system: A rigorous definition. *Journal of Computational Neuroscience*, 2(2):149–162, June 1995. 76

M. Thorne, A. N. Salt, J. E. DeMott, M. M. Henson, O. W. Henson Jr., and S. L. Gewalt. Cochlear fluid space dimensions for six species derived from reconstructions of 3-D magnetic resonance images. *Laryngoscope*, 109:1661–1668, 1999. 11

J. L. van Hemmen. *Theory of Synaptic Plasticity*. Handbook of biological physics(Vol.4), Neuro-infomatics, Neuro modelling, 2001. 44

J. D. Victor. Binless strategies for estimation of information from neural data. *Physical Review E*, 66:051903, 2002. 80

M. Viergever. *Mechanics of the inner ear*. PhD thesis, Delft Technical University, Delft, The Netherlands, 1980. 10, 11

H. L. F. von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. 1863. Published in translation (1954) "On the sensations of tone as a physiological basis for the theory of music", Dover, New York. 2, 53, 76

H. Wang. Information processing in the auditory pathway for automatic speech recognition. Master's thesis, Technische Universität München, 2003. 44

H. Wang and W. Hemmert. Speech coding and information processing by auditory neurons. *Interspeech, Proceedings of the 8th Annual Conference of the International Speech Communication Association*, ISSN 1990-9772:426–429, 2007. 77, 94

H. Wang, M. Holmberg, and W. Hemmert. Auditory information coding by cochlear nucleus onset neurons. In *Proc. IEEE ICASSP'2006*, pages 129–132, Toulouse, France, 2006. 75, 103

H. Wang, D. Gelbart, H. Hirsch, and W. Hemmert. The value of auditory offset adaptation and appropriate acoustic modeling. *Interspeech, Proceedings of the 9th Annual Conference of the International Speech Communication Association*, ISSN 1990-9772: 902–905, 2008. 23, 52, 107

L. A. Westerman and R. L. Smith. Rapid and short-term adaptation in auditory nerve responses. *Hear. Res.*, 15:249–260, 1984. 25

E. J. Williams and S. P. Bacon. Compression estimates using behavioral and otoacoustic emission measures. 201(1–2):44–54, 2005. 17

J. F. Willott and L. S. Bross. Morphology of the octopus cell area of the cochlear nucleus in young and aging c57bl/6j and cba/j mice. 300:61–81, 1990. 38, 44

E. D. Young and M. B. Sachs. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *J. Acoust. Soc. Am.*, 66:1381–1403, 1979. 53, 54, 72, 108

Y. Yu, F. Liu, W. Wang, and T. S. Lee. Optimal synchrony state for maximal information transmission. *NeuroReport*, 15:1605–1610, 2004. 84

X. Zhang and L. H. Carney. Analysis of models for the synapse between the inner hair cell and the auditory nerve. *J. Acoust. Soc. Am.*, 118:1540–53, 2005. 27, 34, 107

X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney. A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression. *J. Acoust. Soc. Am.*, 109:648–670, 2001. 13

Q. Zhu, B. Chen, N. Morgan, and A. Stolcke. On Using MLP Features in LVCSR. In *ICSLP*, 2004. 74