Technische Universität München
Department of Electrical Engineering and Information Technology
Bio-Inspired Information Processing

Master's Thesis

A Comparative Study of Computational Models of Auditory Peripheral System

Nuria Vendrell Llopis

Date of Submission:
September 6, 2010

Supervisors:
Prof. Dr.-Ing. Werner Hemmert
Dipl.-Ing. Marek Rudnicki

Abstract

A deep study about the computational models of the auditory peripheral system from three different research groups: Carney, Meddis and Hemmert, is presented here. The aim is to find out which model fits the data best and which properties of the models are relevant for speech recognition. To get a first approximation, different tests with tones have been performed with seven models. Then we have evaluated the results of these models in the presence of speech. Therefore, two models were studied deeply through an automatic speech recognition (ASR) system, in clean and noisy background and for a diversity of sound levels. The post stimulus time histogram help us to see how the models that improved the offset adaptation present the "dead time". For its part, the synchronization evaluation for tones and modulated signals, have highlighted the better result from the models with offset adaptation. Finally, tuning curves and Q_{10dB} (added to ASR results) on contrary have indicated that the selectivity is not a property needed for speech recognition. Besides the evaluation of the models with ASR have demonstrated the outperforming of models with offset adaptation and the triviality of using cat or human tuning for speech recognition. With this results, we conclude that mostly the model that better fits the data is the one described by Zilany et al. (2009) and the property unquestionable for speech recognition would be a good offset adaptation that offers a better synchronization and a better ASR result. For ASR system it makes no big difference if offset adaptation comes from a shift of the auditory nerve response or from a power law adaptation in the synapse.

Acknowledgments

To my parents, who reminded me in the appropriate time the meaning of the word “courage”.
To Beatriz Macias Ruiz, the S queen.
To Prof. Dr.-Ing. Werner Hemmert, my supervisor Dipl.-Ing. Marek Rudnicki and the rest of BAI team. They made easier the rough and steep climb described by Plato.

Nomenclature

AM Amplitude modulated signals

AN Auditory nerve

AN-IHCS Synapse between uditory nerve and inner hair cell

APS Auditory periphery system

AS Auditory system

ASR Automatic speech recognition

BF Best frequency

BM Basilar membrane

bw Bandwidth

CA Cochlear amplifier

CF Center frequency

DCT Digital cosine transform

DRNL Dual resonance non linear

Eq. Equation

fc Carrier frequency

Fig. Figure

fm Modulation frequency

HSR High spontaneous rate

Hz Hertz

IE Inner ear

IHC Inner hair cell

IHCRP Inner hair cell receptor potential

ISIH Inter-spike interval histogram

LSR Low spontaneous rate

ME Middle ear

MOC Medial olivocochlear system

MSR Medium spontaneous rate

MTF Modulation transfer function

OA Offset adaptation

OE Outer ear

OHC Outer hair cell

PLA Power law adaptation

PSTH Post stimulus time histogram

r Vector strength

SNR Signal to noise ratio

sp/sec Spikes per second

SPL Sound pressure level

Meddis group models' parameters

τ_c Filter time constant to convert BM velocity to cilia displacement

τ_m Calcium current time constant

τ_{CaHSR} Calcium diffusion time constant for HSR fibers

τ_{CaLSR} Calcium diffusion time constant for LSR fibers

τ_{CaMSR} Calcium diffusion time constant for MSR fibers

C_m IHC capacitance

C_{cilia} Cilia/BM coupling gain

Ca^{2+}_{thr} Calcium concentration threshold

CF_{Lin} Center frequency of linear path

CF_{NL} Center frequency of non linear path

E_{Ca} Reversal potential

E_k Potassium equilibrium potential

E_t Endocochlear potential

G^{max}_{Ca} Maximum calcium conductance

G_0 Resting conductance

G_k Potassium conductance

G_{Lin} Gain of the linear path

M_{max} Maximum Free transmitter quanta

R_{pc} Combined resistances

s_0 Displacement sensitivity

s_1 Displacement sensitivity

TW_{delay} Estimate of delay between stimulus and fiber effects

u_0 Displacement offset

u_1 Displacement offset

Carney group models' parameters

C1 Component one of BM-simulation

C2 Component two of BM-simulation

Contents

Chapter 1

Introduction

Since the sixties, several research groups have attempted to model the human ear. The purpose is to establish a clear accurate and precise model of our auditory system (AS), so that we will be able to provide a better understanding of the audition. They are an indispensable tool for observing cochlear processing, these models permit the examination of speech coding in the auditory periphery system (APS) without the need for animal experimentation. As the cochlea is nonlinear, it is a challenge to assign changes in auditory nerve (AN) responses to speech, following each injury to the consequences of the damage, like tuning vs compression. Peripheral models are useful to mimic the cochlear pathology and the responses to noise trauma and accordingly inform hearing aid development. Besides, they are an obligatory requirement for the mimicking of *in vivo* responses in the brain-stem.

Throughout these decades lots of models have been developed. Ones are bio-physiological and try to simulate the underlying physiological processes. Others are phenomenological, they try to simulate, rather than the processes, the result of them. In both cases, each of these models has attempted to provide a further improvement, something that allows us to get closer to the real auditory peripheral system. Not all the optimization have yielded the results that the authors wished. Some have achieved those results, but on the way, have introduced distortions and inconsistencies. The elder models were focused on some particular aspect of the AN response, get a better synchrony, allow two tone suppression, etc, and neglected the rest. Hopefully, recent models cover wider range of responses and pay attention to diverse angles. However, to conduct the future directions on the most worthy way, it is important to compare, rather than the models, the improvements that they develop. Thus, one could focus on upgrades that truly introduce significant changes. This is the aim of this thesis: to elucidate which improvement, among all that have been developed, are worthily for speech.

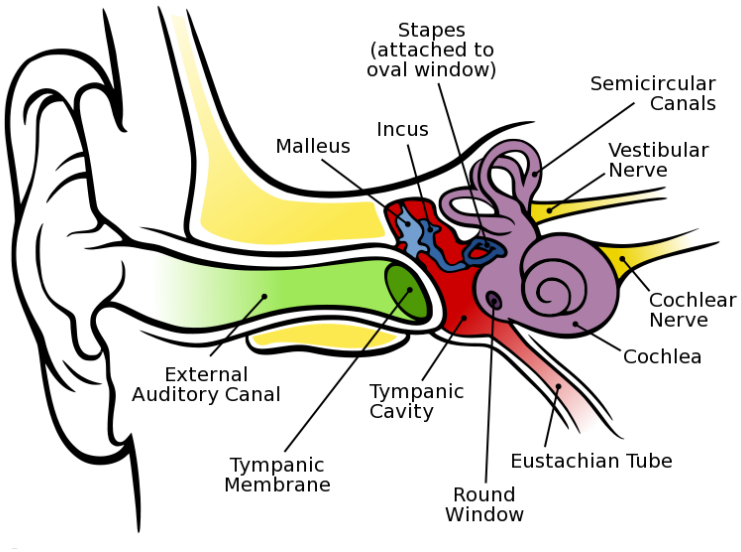

Figure 1.1: The human ear. The eardrum transforms the sounds into mechanical waves, that in turn, pass through the ossicles and arrive to the cochlea. There, it will be converted into spikes. Modified from Chittka and Brockmann (2005)

To go further into the topic, it is important to introduce briefly the physiological basis of the APS and the models we want to check. Thereby, we can understand the complexity of the models, why are they made in the way they are, and the reasons of some improvement or test.

1.1 Physiology of auditory periphery

The AS is responsible for physiological and psychological process of the audition. The sound is transformed in the ear into electrical impulses that are sent to the brain through the auditory nerve. The ear is divided in three:

- Outer ear (OE), that channels the acoustic energy.
- Middle ear (ME), that transforms the acoustic energy into mechanic energy. This energy is then transmitted to the Inner ear.
- Inner ear (IE), where the mechanic energy is transformed into electric impulses.

1.1.1 Outer ear

Its function is to receive sound and to channel it on the eardrum. The shape of the OE amplifies the sound with frequencies between 30-100Hz to around 2-5kHz in human. This bandwidth is also where most the human speech lies.

1.1.2 Middle ear

When the sound arrives to the eardrum, it vibrates and transmits this vibration to the ossicles. These are the smallest bones in the human body and in turn transmit the wave to the fluid in the IE. The ossicles induce an impedance matching between the sound pressure and the fluid waves of the IE. Without this impedance matching, the most of the energy would be reflected and only a small part would be transferred reducing sensitivity. Besides, the ossicles also protect the cochlea from extremely high sound level by uncoupling each other though the action of tensor tympani and stapedius muscles. The peak of efficiency of the ME occurs at 1kHz in humans.

1.1.3 Inner ear

The cochlea, the inner ear part dedicated to audition, is a coiled tube which is composed of three cavities filled with fluids of distinct ionic composition, scala media, scala vestibuli, and scala tympani. Between the last two, we find the basilar membrane (BM), on which the organ of Corti and hair cells reside. When the mechanical energy arrives to the cochlea, a pressure difference appears between the scala vestibuli and tympani, which leads to a deflection of the basilar membrane.

Each area of the basilar membrane vibrates preferentially to a particular sound frequency. For a high frequency the displacement of the BM takes place in the basal zone. That is because, mechanic wave traveling through a liquid is quickly dimmed. Therefore, for low frequencies the mechanic wave travels further and the displacement of the BM happens in the apical zone. See fig. 1.2. For the human being the bandwidth is from 0.1 to 20 kHz.

Lying on the basilar membrane we can find the organ of Corti. At the top are situated the hair cells, and at the bottom, the nerve branches of the auditory nerve. The hair cells collect the vibration of the basilar membrane, which is not uniform but is a function of the resonant frequency of each point of the basilar membrane. Thus, the hair cells generate distinct patterns characteristic for each tone. The human organ of Corti, as mammalian, contains two classes of hair cells.

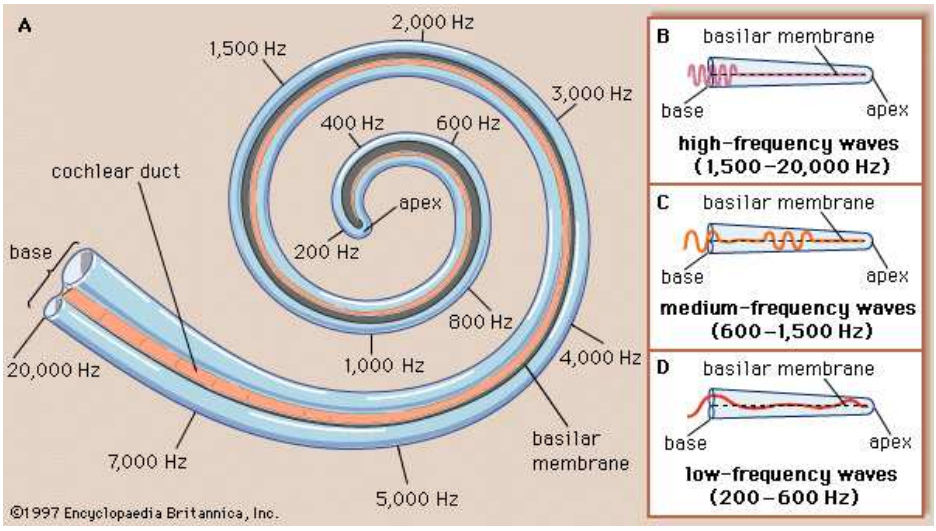

Figure 1.2: Sound frequencies in the BM. A) Different sound frequencies differentially excite different regions of the BM. B), C) and D): varied frequencies and the vibration that they generate.

The outer hair cell (OHC) are responsible for the feedback and are known as "the cochlear amplifier". These cells increase both the amplitude and frequency selectivity through the next mechanism: a deviation of the basilar membrane causes a movement in OHC's hair bundle, that in turn generates a depolarization of the OHC. The depolarization contracts the cells and thus the BM, connected to them, moves in phase. Consequently the cochlea not only responds to the stimulus, but also generates energy by itself. Besides, the OHCs are contacted mainly by efferent nerves, which regulate their electromotility and influence cochlear sensitivity.

The inner hair cell (IHC) is liable for transforming the mechanic wave into electric impulses. Each one of the IHCs has a different center frequency (CF) on which their efficiency is better[1]. This center frequency matches with the vibrate frequency of the BM where the IHC lies. As with the OHC cilia, the IHC cilia displacement causes a depolarization, due to the change in the number of open ion channels. A change in receptor potential appears and causes opening of voltage gated calcium channels. Due to calcium ions entering the cell and accumulating in the vicinity of the synapse, neurotransmitters at the basal end of the cell are released into the synaptic cleft. In the cleft, the

[1]Center frequency and best frequency (BF) have different meanings in the different papers studied here. Usually the BF is the frequency at witch the fiber response is maximum and the CF at which the fiber threshold is the lowest. Nevertheless, it is better to check always the meaning that the authors give to both terms

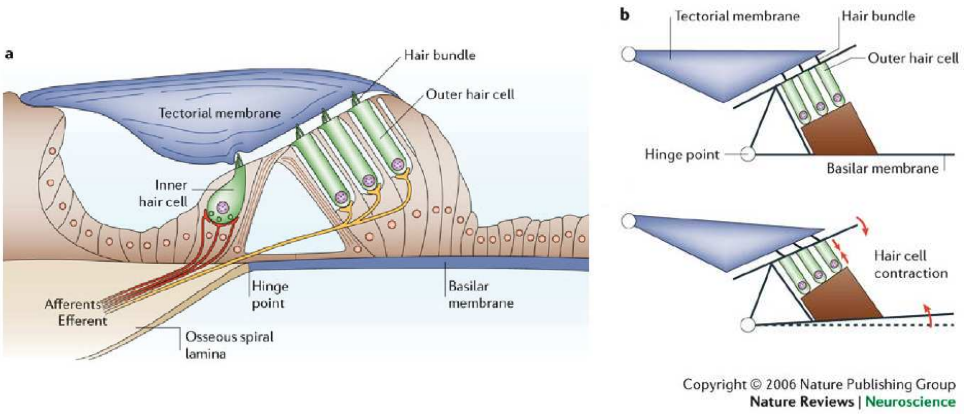

Figure 1.3: A) The organ of Corti. A movement in the BM generates a displacement on the cilia, that in turns produces the depolarization of the hair cells. B) The figure represents the OHC contraction, due to the depolarization. This contraction moves the BM in phase (Fettiplace and Hackney, 2006).

neurotransmitters disperse and some are lost, some are reuptaked, and some bind to the receptors and thus trigger action potentials in the nerve.

1.1.4 Auditory nerve

Auditory nerve fibers connect the hair cells of the cochlea and the cochlear nucleus within the brain-stem. Each hair cell has about ten auditory nerve fibers (AN) with different thresholds connected to it. AN fibers can be of three different types: low (LSR), medium (MSR) or high spontaneous rate (HSR). As the tone amplitude increases, the firing rate of a fiber at CF increases up to saturation. The HSR fibers, which are the most common, saturates rapidly and code intensity changes at low levels. On the other hand, the LSR saturate slowly and code intensity changes at high levels. After firing, an auditory nerve fiber has a refractory period of around 1 ms.

1.2 The models

I have focused on three different research groups, Carney, Meddis and Hemmert, named by their principal researcher. This three groups have been working on the development of an auditory model for many years and have had a big international recognition in their field. They are not the only ones, but from my point of view, they are the most interesting and profitable to

this study, due to their outstanding track record. Besides, I would like to introduce the one from Seneff (1985), because of the importance of her pioneer model.

1.2.1 Models of Carney group

Zhang et al., 2001

The importance of this model on the trajectory of the laboratory is unquestionable. This flagship was the first on a series of models developed by Carney Lab that have taken this one as a basis. Regarding the model, the aims are to provide a more accurate and quantitative description of the responses of the AN-HSR to complex sounds and to find the way to understand the several nonlinear response properties. It also provides a tool to comprehend the AN population response through the study of these nonlinear encoding. All this, trying to keep the model as simple as possible. The BM was mimicked by two paths. The control path, through the level dependent gain, bandwidth and phase properties, manages the signal path filter.

Thanks to the control of bandwidth, the wide frequency range of two-tone suppression was included. The control path has a level dependent filter with a frequency higher than the signal path filter. As the result of that, the researchers could also include in the model the asymmetrical growth of suppression above and below the characteristic frequency and suppression tuning curve frequency offset.

The nonlinear tuning, without an increase in the complexity from previous models (Carney, 1993), is much more accurate. On the other hand the IHC section consists of a logarithmic saturating function followed by a seventh-order low-pass filter. The IHC-AN synapse model is a time-varying model only for HSR.

Heinz et al., 2001

In this paper, they develop a model based on Zhang et al. (2001) to simulate normal and impaired human peripheral auditory. For the first time in Carney lab, three different types of AN: HSR, MSR, LSR are included. Besides, the model offers five different control path to evaluate different AN's and some parameters have been changed. However, the most important difference from the previous model is the fact that they used a human cochlear map. As a result of this, the model was used subsequently in Tan and Carney (2005) to study the encoding of vowel-like sounds.

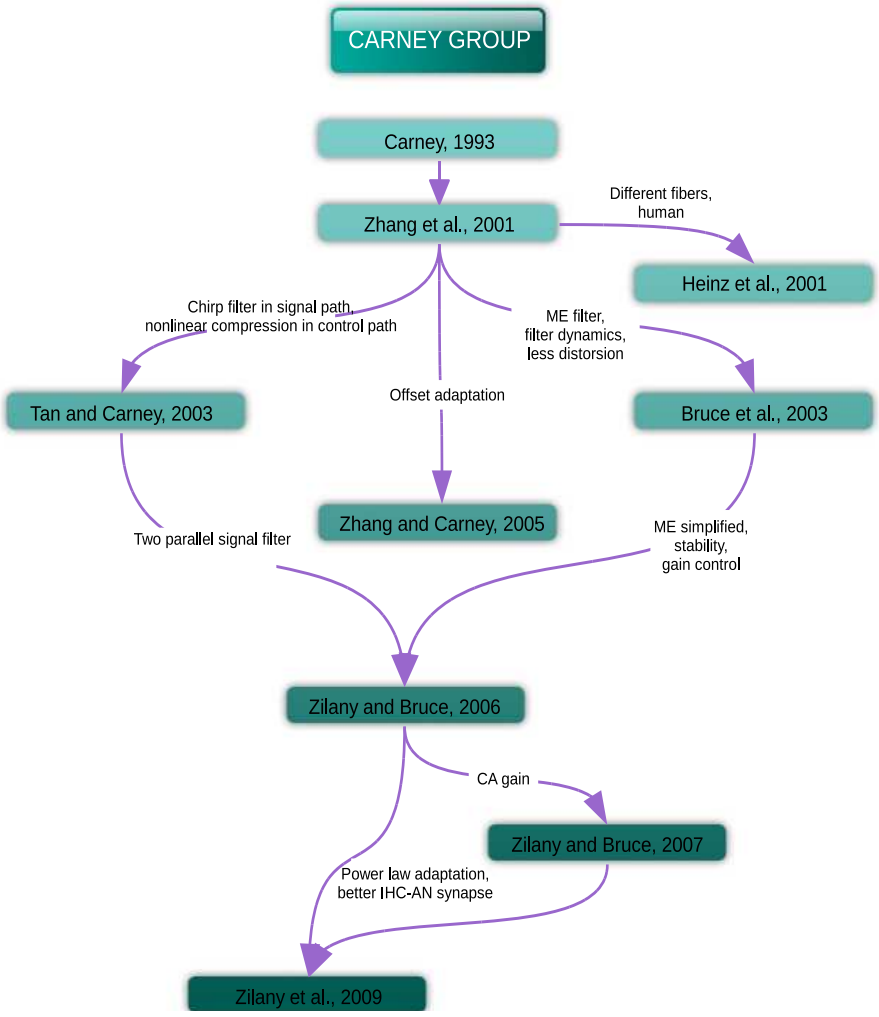

Figure 1.4: Carney group's models

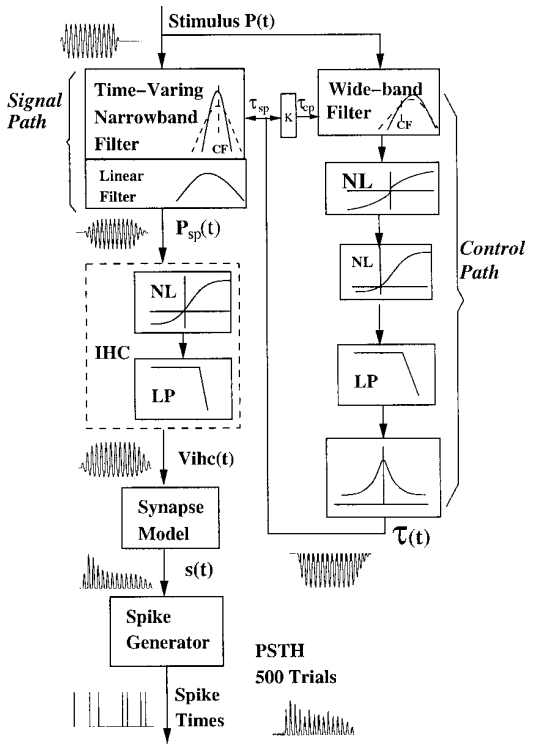

Figure 1.5: Block diagram of the Zhang et al. (2001) AN model

Bruce et al., 2003

With the intention of investigating how the IHC impairment affects on the broadened tuning on responses to speech, the authors modified the model of Zhang et al. (2001). Specificallys they modified the OHC and IHC sections to simulate an impaired AN and observe how the threshold and bandwidth changes and how it affects on model responses. They also include a ME filter that does not exist before and modify the control path to improve the filter dynamics. This new control path avoids the distortion products created by the compression of the signal. Thereby a dynamic filter is used instead of an static nonlinearity. This model simulates only the low and moderate level responses in normal and impaired ears. We had to wait to Zilany and Bruce (2006) to simulate the high level responses.

Tan and Carney, 2003

The aim of this research was the study of peripheral auditory processing crossing arbitrary sounds inputs through the model. It attempts to simulate the responses of AN fibers in cat with more complete response features than previous ones. Besides, they emphasize on mimicking the level-independent frequency glide and its implication on stimulus encoding. Definitely, they managed to include in the model the instantaneous frequency glide and compressive nonlinearity. The model inherits the IHC section of Zhang et al. (2001) and possesses its own signal and control path of the BM and a filter section for the ME. The signal path corresponds to a varying band-pass filter and the control path with a nonlinear compression. The model has been used by Tan and Carney (2005) for the same reason as in section 1.2.1 and in Tan and Carney (2006) to understand how the AS extract speech signals in presence of noise.

Zhang and Carney, 2005

Here we can find a comparison between the models of Sumner et al. (2002, 2003a) and Zhang et al. (2001). This comparison is made to study "*the effects of adaptation characteristics on model parameters and of model parameters on adaptation characteristics*" (Zhang and Carney, 2005). Moreover, they use both model to learn about the offset adaptation (OA) and how the models structures limit it. Then they modified both with an improved, more realistic OA and ameliorated the modulation gain of model AN fiber responses to modulated stimuli. However, on the way they introduce an unexpected variation in the average rate with modulation frequency and

an undesirable unrealistic steady-state rates of LSR fibers to tones at high sound levels (Zilany et al., 2009).

Zilany and Bruce, 2006

The aim of this study is to find a better description of the AN response properties for a large range of CF's, simulating low, moderate and high level responses in cats. The input of the models is intended to be complex or simple stimuli spanning the sound pressure level (SPL) dynamic range of hearing. The idea is to suggest and test new strategies for hearing-aid signal processing through accurate high level models, since hearing aids amplify signals to compensate for hearing loss. To introduce the high level modeling, it has been necessary to change the previous model. It particularly differs from the model presented for Bruce et al. (2003) in the use of two modes of BM excitation. Each of the two parallel models has its own transduction functions, that later are summed and passed through the IHC low-pass filter and the IHC-AN synapse of Zhang et al. (2001) with negligible modifications. In addition, the ME section of the model is simplified from Bruce et al. (2003) to ensure the stability and the control path remains the same, except for some parameters.

Zilany and Bruce, 2007

The model of Zilany and Bruce (2006) has been modified to introduce an slight adjustment in the cochlear amplifier (CA) gain to improve the model prediction of the vowel data. The CA gain in the CF range must be enlarged to fit the model result. This modification has a paltry effect over the rest of the properties of the model response. It is designed to accurately deliver the response of AN fibers when the input is a steady-state vowel.

Zilany et al., 2009

This study tries to shed light on the mechanism that gives rise to synaptic adaptation. A previous research work (Zhang and Carney, 2005) has attempted to reproduce the onset and OA with the problems we have seen in section 1.2.1. Introducing the power law adaptation (PLA) to the previous model of Zilany and Bruce (2006, 2007), model responses were compared to physiological data. The PLA synapse model improves the AN responses remarkably at stimulus offset and also recoveries after stimulus offset. It also increases synchrony to pure tones.

1.2.2 Models of Meddis group

Meddis et al., 2001

Here, an algorithm to simulate the basilar membrane is formulated. It is implemented as a dual resonance non linear (DRNL) filter. The DRNL is divided in two parallel paths, linear and nonlinear, with different CF. The nonlinear path includes a compress of the input signal when it exceeds a threshold level. This input to the algorithm is stapes motion and the result of the model is the sum of the two paths and represent basilar membrane motion.

Sumner et al., 2002

Meddis group has develop their models as a connection of different components: middle ear, a simple pass-band filter; basilar membrane, a nonlinear signal-processing algorithm; inner hair cell receptor potential (IHCRP) and inner hair cell synapse (AN-IHCS), both a complex algorithm. In this paper, an algorithm to mimic the IHCRP and the AN-IHCS is presented. The BM component is taken and briefly modified from a previous publication (Meddis et al., 2001). However, Sumner et al. (2002) emphasize the development of the IHC. They theorize that the rate of neurotransmitter release is controlled and doubtlessly modeled by the presynaptic calcium current. This release into the cleft determines the action potential rate of the AN.

The model is able to reproduce the rate intensity response of HSR, MSR or LSR fibers. This fairing rate variation could be managed by the maximum calcium conductance. Phase-locking characteristics, relative refractory effects, mean-to-variance ratio and discharge history effects could be also reproduce by the model.

Sumner et al., 2003a

In this work, the adaptation of the model presented in Sumner et al. (2002) is depicted. The characteristics of this adaptation depends on the type of fiber that are determined by the number of calcium channels near the synapse. Only some changes in the parameters of the model to simulate different neurons are performed.

Sumner et al., 2003b

Working with almost the same model as Sumner et al. (2002, 2003a) they attempt to reproduce a wide range of responses to auditory stimulation. The

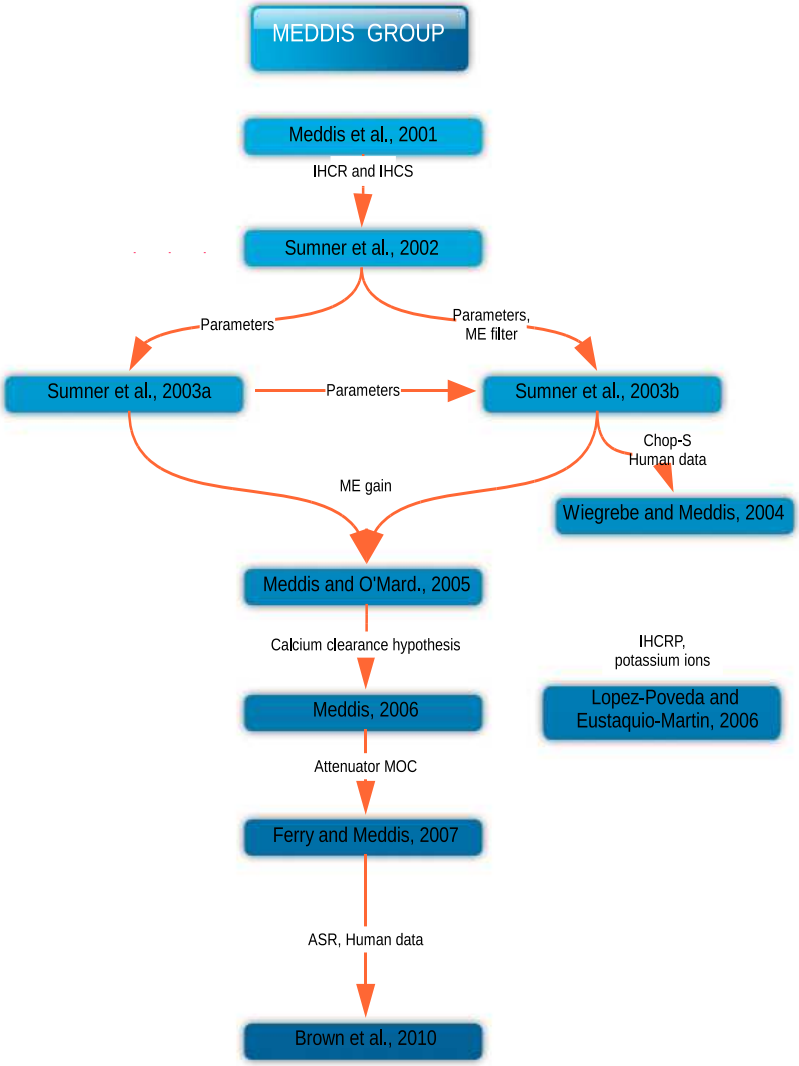

Figure 1.6: Meddis group's models

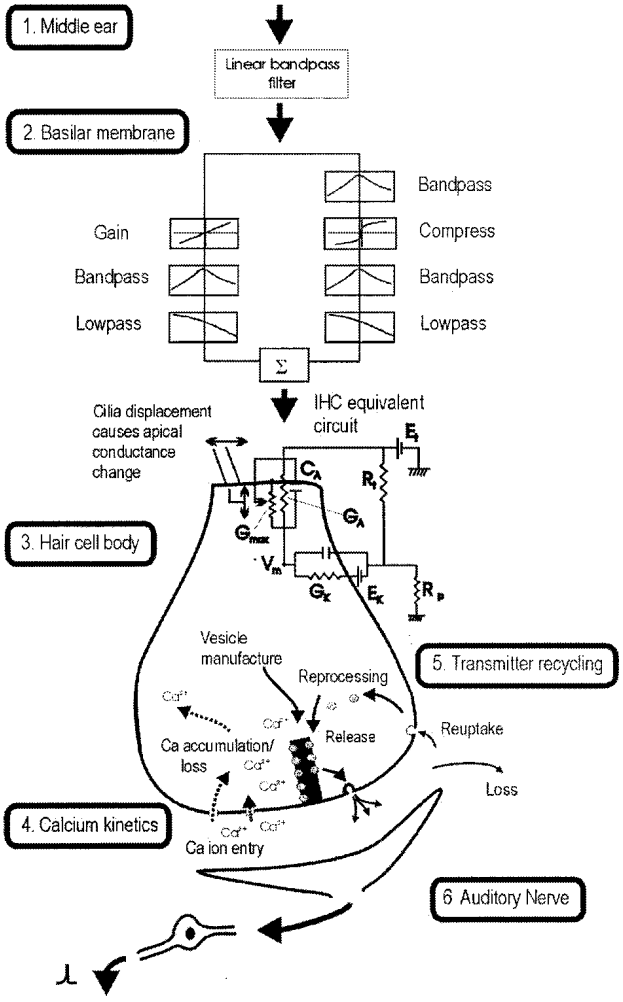

Figure 1.7: Block diagram of the Sumner et al. (2002) auditory model

idea is to use it, as an input to a larger model which mimics the auditory processing in the brain-stem. For that reason they change the ME filter to a cascade of two linear band pass Butterworth filters and some parameters of the rest of the components.

Wiegrebe and Meddis, 2004

The role of sustained chopper (Chop-S) in the extraction of pitch, through the application of an autocorrelation to the temporal discharge patterns of AN fibers is studied here. To rise their goal, this paper fits the parameters of previous models (Sumner et al., 2002, 2003a,b) to human data. Thus the ME filter, the DRNL (from (Lopez-Poveda and Meddis, 2001)), the IHCRP and IHC parameters are changed.

Meddis and O'Mard, 2005

The aim of this study is to distinguish the role of the AN adaptation in forward-masking effects. The model includes a coincidence-detection mechanism for making threshold decisions depending on the fire of an AN fibers group. The results of the simulation suggest that "*poststimulatory reductions in AN activity can make a substantial contribution to the raised threshold*" (Meddis and O'Mard, 2005)

The same model as Sumner et al. (2002, 2003a,b) is used, except for the parameters that are taken from Sumner et al. (2003b) and the scaling factor in the ME simulation. This last parameter is changed to increase the ME gain.

Meddis, 2006

In this paper the authors introduce a new version of AN-IHCS. Since they wanted to demonstrate that "*auditory-nerve fiber spikes can be predicted to occur when the running integral of stimulus pressure reaches some critical value*"(Meddis, 2006), they examined two different "presynaptic calcium" to explain this effect. The first one, the "Calcium influx model", comes from previous publications (Sumner et al., 2002, 2003a,b; Meddis and O'Mard, 2005). The second one, the "Calcium clearance hypothesis", uses the same model but with changes in the parameters that allow us to simplify the equations and make the relationship between calcium levels and transmitter release clearer.

Lopez-Poveda and Eustaquio-Martin, 2006

A new version of the IHCRP component is delivered in this work . Unlike the IHCRP component from previous models, this one is based on the contribution of the basolateral potassium currents on the IHC nonlinear input/output transfer characteristics. The component attributes the responsibility of the IHC membrane conductance to potassium ions only, ignoring sodium and chlorine ions.

Ferry and Meddis, 2007

In this model, they adapt the previous computer component of BM (Meddis et al., 2001) to simulate the effect of the medial olivocochlear system (MOC) and his role in auditory processing of complex sound. The purpose of olivocochlear bundle "*appears to be the regulation of activity in the AN through modification of OHC electrical and mechanical properties and, more directly, through post-synaptic contacts on the AN itself*" (Ferry and Meddis, 2007). The MOC, one of the two parts of the olivocochlear bundle, influences suppressively, via efferents, on the response of the BM (Dallos, 1992). Furthermore, this suppression enhances the response in adverse background as noise (Kawase and Liberman, 1993)

Brown et al., 2010

The newest model of Meddis group develops a study to ascertain the role of the MOC in speech recognition with broadband noise. As said before (in section 1.2.2), the efferent system can reduce the response to the continuous noise by reducing the adaptation. This may be the reason of the speech intelligibility of normal human listeners in the presence of background noise. Thereby, an auditory model that takes into account this efferent processing could be a noise-robust front-end for an automatic speech recognition (ASR) system in adverse acoustic conditions. This hypothesis is supported by the fact that noise, if any, is the dominant influence on the amount of efferent activity (Liberman, 1988) and with the results of the present paper. It has become apparent that, when noise is present, speech recognition accuracy is improved by simulating the attenuation of the BM response. Furthermore, the optimum efferent attenuation is proportional to the noise level.

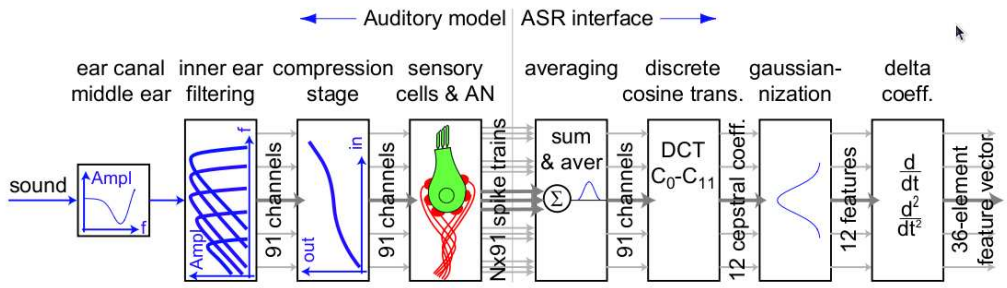

Figure 1.8: Schematic of the Holmberg et al. (2007) auditory model

1.2.3 Models of Hemmert group

Holmberg et al., 2007

The authors develop a model that mimic significant characteristics of auditory nerve spike trains. They attempt to observe the rate-place coding strategy and an interval-based strategy to speech encoding. They explore the result of introducing the information that is coded in the auditory pathway into an ASR. Thereby, the importance of temporal coding result evidenced. The model is based on Holmberg (2007). The ME is mimicked as a high-pass function and the BM is simulated with a computational wave-digital filter. In contrast, the IHC model is extracted and modified from Sumner et al. (2002). To simplify the model, the pools are quantized instead of continuous vesicle pools. In the following paper, its degradation of the synchronization index above 1kHz will be solved with the introduction of an OA process.

Wang et al., 2008

In that study an OA procedure is introduced to the model of Holmberg et al. (2007). This OA model comes from Zhang and Carney (2005). The purpose is to make the onset neurons located in the auditory brain-stem responsive to the frequency bandwidth above 3kHz. As a result of the modifications, they get the same onset adaptation but a better and realistic OA, more precise phase locking to amplitude modulated stimuli and improved ASR results.

1.2.4 Seneff

Due to the difficulty in the eighties of designing a worthy computational speech recognition, the researchers were "*deterred from designing a speech analysis system that is motivated by the human auditory response mechanism*". Swimming against the tide, the thesis of Seneff (1985) developed a

way to detect synchrony in the response to predictable periodicites. These might be able to emphasize peaks in the spectrum. Thereby, its application to the spectral analysis and reckoning to the fundamental voice frequency would be delivered. For that reason, Seneff designed a model that takes the incoming speech signal and to process it, uses a system which mimics the auditory peripheral system. Then, to point up the spectral concept that are significant of the recognition speech, she employed a synchrony measure.

The implementation of the Seneff's model was developed by Slaney (1998) who through Matlab offers a collection of tools to reproduce the APS.

1.3 Objectives

The most important goal of this project is first to find out which model fits better with the data and which property are relevant for speech recognition. This information could help the research groups that perform this kind of models to clarify the next steps to follow. To achieve this goal some sub-goals are required, such as:

- Implement an interface for all the models which provides with an equal platform with the same input (simple or complex audio signals) and output signals (afferent responses of the AN fibers).
- Develop different tests over simple signals (as explained in chapter 3) to get an overview about the properties of the models.
- Carry out a test with complex signals. The output of the models (afferent responses) is the input of an ASR. The output of the ASR will be a percentage of recognition to discover how accurate are the models' responses. This test will be made for different stimulus signals, with and without noise and for different signal levels.
- A large study with the information compiled to culminate with our study.

Once the project is completed and all models have been studied among all the tests proposed, we will be able to select the best improvements from the results. Contrasting the percentage of speech recognition over a variety of circumstances, countable differences will be pointed out among the models. Those will be the primary method of decision. The simple tone stimulation test will help then to explain these differences found with the ASR. Therefore, we will be able, not only to offer the best improvements, but also the reason why.

Chapter 2

Methods

The models of the APS are similar in its block system. They model every stage of the audition into blocks that can be usually interchanged by new ones. This development with blocks allows the improvement of every part of the APS independently. The first model block is normally the ME, represented as a filter. Then the BM is simulated as a nonlinear function of frequency. For the Meddis group, this BM has two paths of filters, linear and nonlinear. For Carney group the BM is also divided in two paths but one path controls the other. For Hemmert group, the BM is a cascade of resonators. The IHC, a nonlinear function, is the next step followed by the AN-IHCS, both differ mostly from one group to other. But not only they differ among research groups. The improvements introduced in every new model have changed one or more of this blocks mentioned and the way they develop it. In this chapter I will explain this differences in implementation and development. Besides, in this chapter, I will also mention the particularities of the test (simple and complex signals) I performed and how is implemented the ASR block is implemented to get the amount of recognition percentage.

2.1 Interface

In this thesis I have worked with a motley group of models. Most of them have their codes available in their website, although others have been requested personally. Sometimes the code was not completed or the publication had some mistakes, thus, a new implementation and test of the models have been compulsory. Besides, this code is developed under different platforms, such as Matlab, C or Python. There has also been a need to present a common interface for all the models. Thereby, the input of the system and the output

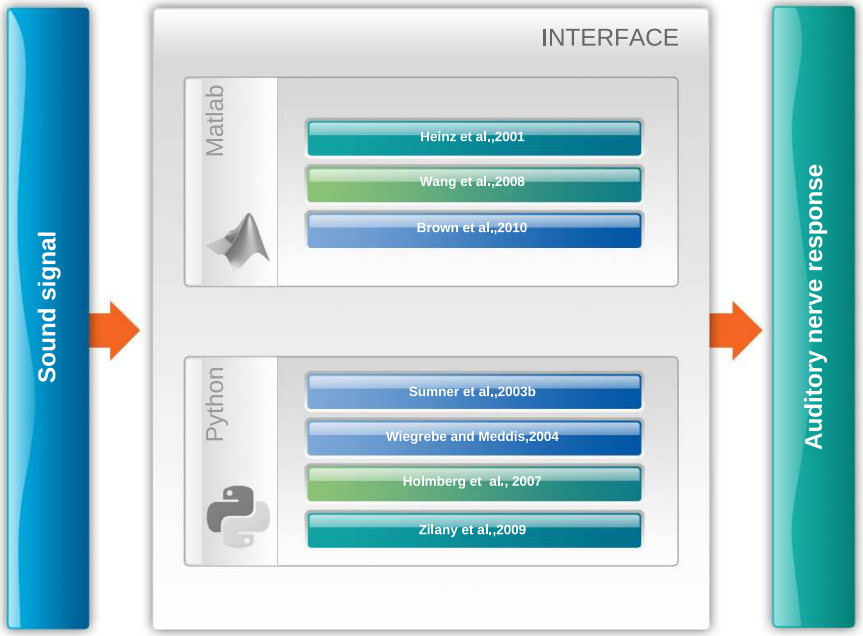

Figure 2.1: Interface: the input and output of the system are the same for all models.

are independent of the model used, and only depend on the simulation that is running. The tests of the section 2.4 have been developed in Python and Matlab.

2.2 Models

Seven different models have been chosen to be analyzed, two models from Carney group, three from Meddis group and two from Hemmert group. The implementation and the parameters used are described below.

Model of Heinz et al. (2001)

This model is the same as developed in Zhang et al. (2001) fig. 1.5, with some modifications. The input to the model is a sound pressure (μPa) signal. The middle ear is dismissed here and the BM is modeled as a nonlinear filter with two paths, signal and control. The signal path is a nonlinear, third order, time-varying narrowband filter and a linear, first-order broadband filter in

cascade. The control path output varies the gain and bandwidth of the nonlinear signal filter. Besides, the control path has a nonlinear wideband filter followed by a nonlinear saturation that represents the properties of the transduction of OHC. In this model, unlike in Zhang et al. (2001), the human ear is pretended to be modeled. Thus, the human cochlear map is used. After this, a third-order low-pass filter fits the dynamics of the cochlear amplifier. The IHC is described as a logarithmic compressive function and a seventh order, low-pass filter. This low-pass filter was changed from the previous model to satisfy the inclusion of different fibers. The parameters used to get these fibers are, HSR = 60 sp/sec, MSR = 10 sp/sec, LSR = 1sp/sec. The nonlinear AN synapse is a simplification of the one explained in Carney (1993), a time-varying three-store diffusion model. The synapse output of Heinz et al. (2001) is a probability function. However, a spike generator is also offered.

Model of Sumner et al. (2003b)

The structure of the model is inherited from the model of Sumner et al. (2002) fig 1.7. The input to the model is the instantaneous pressure waveform of the stimulus (μPa). The ME filter consisted in a cascade of two linear band-pass Butterworth filters, a second order filter and a third order filter with [4,25]kHz and [0.7,30]kHz respectively (Fig. 2.3). Both have unity gain in the pass-band, although a variable gain G_{me} and a scaled factor ($1.4*10^{-10}$ m/s/μPa) are also introduced here. Then, the signal passes through a DRNL filter consisting of two parallel pathways (one linear, one nonlinear) as in fig 2.2. The linear pathway is a gain, a gammatone filter and a low-pass filter. The nonlinear pathway is a gammatone filter followed by a compression function, a second gammatone filter and a low-pass filter. The parameters, a, b, the bandwidth of linear (BW_{lin}) and nonlinear pathways (BW_{nlin}), the gain of the linear filter (G_{lin}) and the CF of the linear filter (CF_{lin}) vary linearly on a log-log scale with the following expression:

$$Parameter(CF) = 10^{p_0 \, + \, m \, * \, log_{10}(CF)} \tag{2.1}$$

where $\mathbf{p}_0$ is the parameter value at a BF of 1 Hz and $\mathbf{m}$ is the slope of the parameter with frequency. This parameters are taken from (Sumner et al., 2003b, Table I). Both outputs summed, linear and nonlinear, are the input for the next stage, the IHCRP, a biophisical model of the cilia transduction and receptor potential response, widely explained on Sumner et al. (2002). The parameters G_{Ca}^{max} and Ca_{thr}^{2+} for control of calcium levels are the responsible for the different types of fibers and change from Sumner et al. (2002) as shown also in (Sumner et al., 2003b, Table I). The last stage, the synapse,

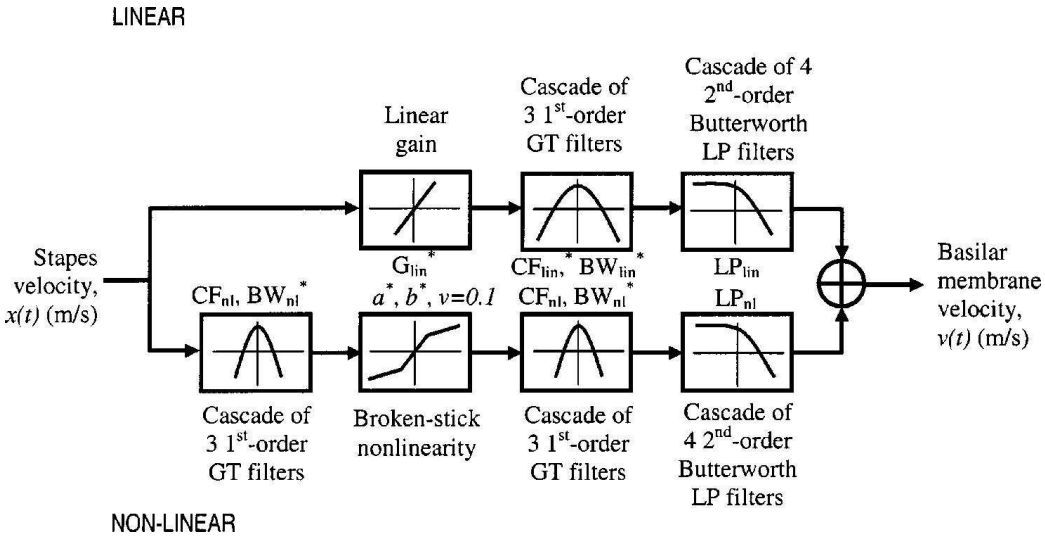

Figure 2.2: DRNL filter architecture.(Sumner et al., 2003b) The asterisk are parameters that have change from Sumner et al. (2002)

is simulated by a uniform random process. The parameters of this part and the rest of parameters not specified here, remain the same as in previous publications Sumner et al. (2002, 2003a).

DRNL filter parameters (CF_{NL})	**sum2003b**		**wie2004**	
	p0	**m**	**p0**	**m**
Bandwidth of non linear path BW_{NL}	0.8	0.58	-0.032	0.774
Compression parameter, a	1.87	0.45	1.4	0.82
Compression parameter, b	-5.65	0.875	1.62	-0.82
Center frequency of linear path CF_{Lin}	0.339	0.895	0.037	0.79
Bandwidth of linear path BW_{Lin}	1.3	0.53	0.037	0.79
Linear path gain G_{Lin}	5.68	-0.97	4.2	-0.48
Compression exponent, v	0.1		0.25	

Table 2.1: Table of recalculated coefficients **m** for computing parameters of the DRNL filters as a function of CF_{NL}. The column sum2003 represents the values given in Sumner et al. (2003b), and the column wie2004 represents Wiegrebe and Meddis (2004) values

Model of Wiegrebe and Meddis (2004)

This model uses the same model as explained in Sumner et al. (2002, 2003a,b), fig 1.7, with the same structure explained above, but with some changes in

IHC parameters	**sum2003b**			**wie2004**
	HSR	**MSR**	**LSR**	**HSR**
Ca^{2+}_{thr}	0	$3.35*10^{-14}$	$1.4*10^{-11}$	$4.48*10^{-11}$
M_{max}	10	10	10	10
G^{max}_{Ca}(S)	$7.2*10^{-9}$	$2.4*10^{-9}$	$1.6*10^{-9}$	$8*10^{-9}$

Table 2.2: Table of IHC parameters, for three different types of fibers in sum2003 (Sumner et al., 2003b) and one type in wie2004 (Wiegrebe and Meddis, 2004)

the parameters explained below. First the ME filter is a four parallel second order band-pass filters: a) -12 dB gain and a band-pass of [0.1,1.3]kHz; b) 1.5 dB gain and a band-pass of [0.35,6.5]kHz; c) 5 dB gain and a band-pass of [1.8,5.2]kHz and d)-11 dB gain and [7.5,9.9]kHz of band-pass (Fig. 2.3). Besides, the stapes scalar factor is $3 * 10^{-10}$ m/s/μPa. The parameters values of the DRNL have been taken from (Lopez-Poveda and Meddis, 2001, table III), as is explained in the publication, opposite to the rest of the parameters that have been taken from the appendix of the publication, such as, the total capacitance ($C_m = 15 * 10^{-12}$ F), the cilia/BM time constant ($\tau_c = 2.13 * 10^{-4}$ s) and the cilia/BM coupling gain ($C_{cilia} = 0$ dB) that vary in the IHCRP. The modifications in the IHC parameters are presented in table 2.2 and lastly the reprocessing rate, parameter of the AN-IHCS, have been modified to x= 90 s^{-1}. Besides the maximum CF that can be used in this model is 6kHz.

Model of Holmberg et al. (2007)

This model is divided in different parts as shown in fig. 1.8. The middle ear and eardrum transfer function consists of a high-pass first order filter of f_c=1kHz and an IIR tenth order filter (fig.2.3) respectively. Then, the model has a transmission-line model consisting of digital filters equivalent to a series of second order resonators, the electrical equivalent circuit of the BM. To simulate the next compression stage and the nonlinear amplification, each of these resonators are followed by another four resonators that are, in turn, modelated as digital filters. The IHCRP and IHCS are adapted from the one presented in Sumner et al. (2002) and from there, the parameters to run the model are taken.

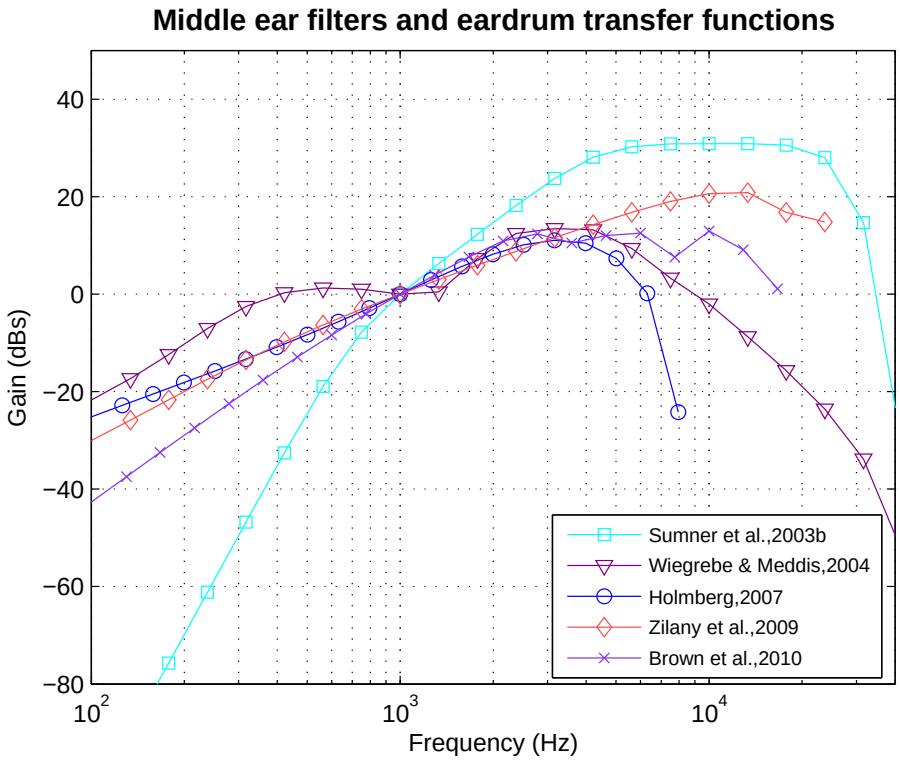

Figure 2.3: The middle ear filters and eardrum functions (when present) of the different models normalized at the same gain (0 dB) at 1kHz. The ME presented in Wang et al. (2008), has the same shape as the one of Holmberg et al. (2007). Besides, the model described by Heinz et al. (2001) has no middle ear filter.

Model of Wang et al. (2008)

The main improvement of the improved model of Hemmert group is the inclusion of an OA. The pool model described by Holmberg et al. (2007) uses the same function for offset and onset adaptation, the model recovers immediately and thus, the model can not reproduce the "dead time" after the end of the stimulus. The present model uses an OA based on Zhang and Carney (2005) that adds a shift value to the synaptic output for negative values. This values are, in turn, set to 0 that represent the "dead time". Thereby, the model will reproduce accurately the onset and offset adaptation. This model calculates the response of the auditory nerves to an amount of different frequency channels, from 49Hz to 14kHz (using the frequency map explained below). For that reason, is needed to select the correct channel from the output of the model when a simple tone test is pretended.

Model of Zilany et al. (2009)

The newest Carney model has the same structure as the previous models in their group with some differences. First, the middle ear model changed in a second revision of code from the original published released[1]. Now it is formed by 3 IIR filters with the following formulas.

$$ME_1(z) = 0.997 * \frac{1 - z^{-1}}{1 - 0.996 * z^{-1}} \tag{2.2}$$

$$ME_2(z) = 2.3 * 10^{-11} * \frac{4.1 * 10^{10} - 7.8 * 10^{10} * z^{-1} + 4.0 * 10^{10} * z^{-2}}{1 - 7.9 * 10^{10} * z^{-1} + 3.7 * 10^{10} * z^{-2}} \tag{2.3}$$

$$ME_3(z) = 2.4 * 10^{-11} * \frac{1.1 * 10^{11} - 1.4 * 10^{8} * z^{-1} - 1.1 * 10^{11} * z^{-2}}{1 - 7.8 * 10^{10} * z^{-1} + 3.9 * 10^{10} * z^{-2}} \tag{2.4}$$

Second, it inherits the two modes of BM excitation explained in Zilany and Bruce (2006). This two different and parallel filters component one (C1) and component two (C2) generate the different parallel modes of the signal path. C1 is designed to get the low and medium level response, is implemented as a chirp filter (Tan and Carney, 2003) and is able to reproduce frequency glides and BF shifts. This C1 has been adapted to produce more realistic frequency tuning curves, more realistic level-independent frequency glides in the impulse response and an improved simulation of the phase properties of the Cochlea. C2 manages the high levels response, is broadly tuned, is resistant to trauma and is shifted in phase 180^o to C1. Consists of a filter

[1]The values were taken by printing them in the process of the analysis of a 1kHz signal input in the model

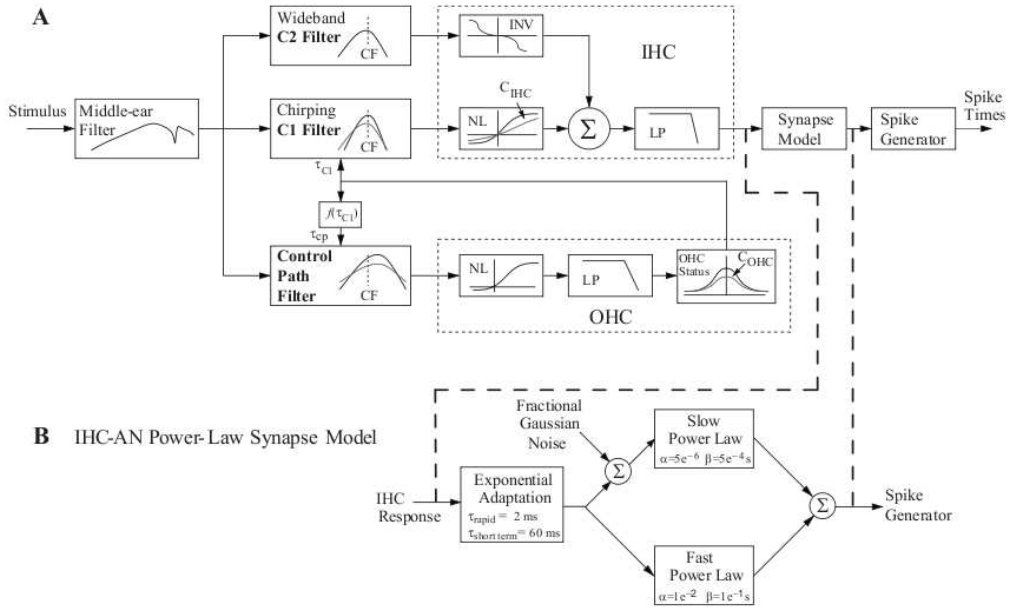

Figure 2.4: Schematic diagram of Zilany et al. (2009) model. To A corresponds the basic Carney group model until Zilany and Bruce (2006), with the single path (C1), parallel path (C2) and control path, followed by the IHC section and in turn, the synapse. B represents the improvement of the model, the PLA model and its slow and fast paths.

based on Kiang's two-factor cancellation hypothesis (Kiang, 1990), the OHC-impaired version of the C1 filter. Besides the IHC low pass filter was reduced to 3 kHz to adjust the maximum synchronized response of AN due to the increase on synchrony to pure tones. Lately, the most important change is the inclusion of power-law functions. Following the exponential adaptation, the model presents two paths of PLA, slow and fast power-law adapting components that will be summed after (in the slow path, the fractional Gaussian noise is added to model the distribution of spontaneous rate). The PLA function represents the convolution of the power law kernel with its previous response. This kernel can be approximated by sixth and tenth order IIR filters respectively. This approximation has been made to run the ASR due to the computational expensiveness of the power law function. Besides, for ASR tests, the model has used his own frequency map (instead the one presented below), because it can not work in frequencies under 80Hz.

Model of Brown et al. (2010)

This model is divided in two stages. The first stage does not differ greatly from the one presented in Ferry and Meddis (2007), the parameters are changed to simulate human hearing instead of guinea-pig, but the structure of the model remains the same. To get an anti-masking effect (the best improvement of this models), an attenuator proportional to the amount of MOC activity was introduced to the model of Meddis (2006). It is situated at the beginning of the nonlinear path of the DRNL module used to simulate the response of the BM. Thereby the model was able to simulate the effect of the MOC independently of the fiber type. The second stage is completely new and corresponds with an ASR system that uses statistical word model to convert AN firing pattern into word sequence.

Develop to work with ASR, this model has nine different modules such as, acoustic stimulus, stapes, BM motion, IHC stereocilia, IHCRP, IHC transmitter release, HSR-LSR auditory nerve action potentials and two moduls of brainstem response: the first represent the chopper cell and the second, the chopper unit. For the tests I used all the moduls but I got the output from the seventh modul, that gives the AN spike response with efferent effects. The middle ear filter shown in fig. 2.3 is implemented as a high-pass first order filter followed by a parallel structure of three different first order linear Butterworth filters with [1.9,4.2]kHz, [4.5,6.3]kHz and [8,12]kHz band-pass respectively. The BM is modeled as a DRNL filter with a third order gammatone filter in the linear and nonlinear path. The parameters of the model where chosen as in table 2.3 from the parameters files offered by request by R. Meddis. The seventh model AN-IHCS output can be "probability" or "spikes". However, the spikes output has been used in all the tests due to the efferent system used.

2.3 Frequency map

The frequencies that have been used to perform the tests fit the human frequency map described in Holmberg (2007). That is a frequency array that starts at 49Hz and finishes at 14197Hz following a logarithmic formula as:

$$f_{CF}(x) \;=\; (10^{1.8 \,*\, (L_{BM}-x)/L_{BM}} - 0.8) \;*\; 238Hz \tag{2.5}$$

where $\mathbf{L_{BM}}$ is the total length of the cochlea ($L_{BM} = 0.035$m) and $\mathbf{x}$ is the position on the cochlea.

DRNL

stapes factor	$6*10^{-8}$	linear path gain factor	$3*10^{5}$
a, nonlinear path gain	$4.5*10^{5}$	b, sets compresion (m/s)	$18*10^{-5}$
	10^{5}		$50*10^{-5}$
compression parameters	0.2		

IHC stereocilia

τ_c	0.0003	C_{cilia}	0.1
G_{Ca}^{max}(S)	$8*10^{-9}$	G_0	$1.97*10^{-9}$
u_0	$40*10^{-9}$	u_1	$7*10^{-9}$
s_0	$1*10^{-7}$	s_1	$20*10^{-9}$

IHCRP

C_m (F)	$16*10^{-12}$	E_t (V)	0.09
G_k (S)	$1.8*10^{-8}$	E_k	-0.0705
R_{pc}	0.04		

IHCpreSynapse

E_{Ca}	0.066	z	$2*10^{42}$
β_{CA}	400	γ_{CA}	130
τ_{CaHSR}	$2.1*10^{-4}$	τ_{CaMSR}	$1*10^{-4}$
τ_{CaLSR}	$0.6*10^{-4}$	τ_m	0.0001

AN-IHCS

M_{max}	12	y, replacement rate	3
l, loss rate (s^{-1})	2580	replenishment rate (s^{-1})	30
r, reuptake rate (s^{-1})	6580	refractory period	0.00075
TW_{delay}	0.004		

Table 2.3: Table of Brown et al. (2010) parameters

2.4 Simple tone stimulation

Post-stimulus-time histogram (PSTH)

To visualize the difference in timing between the models, the PSTH has been conducted. The PSTH is an histogram of the times at which neurons fire. We used a tone burst of 25 ms with 1 ms of rise and fall time, the same as the signal used for the data showed, and a frequency of 4.3kHz. Due to the peak-to-sustained discharge rate increase with SPL Carney (1993), the sound level was fixed to 30 dB SPL over the threshold at this frequency. The fibers tested were 1000 HSR fibers with a CF of 4.3kHz. The resolution used was fit to a bin size of 0.5 ms.

How was got the rate intensity

The three different types of AN fibers, HSR, MSR and LSR, have been tested here among a large range of intensities. The amount of fibers of each type was 250 with a high CF. This CF changes between models: while for the model of Sumner et al. (2003b) and Zilany et al. (2009) was 14kHz for the model of Heinz et al. (2001), Wiegrebe and Meddis (2004),Holmberg et al. (2007), Wang et al. (2008) and Brown et al. (2010) was 4.3kHz. The reason is that, the two first try to simulate guinea-pig and cat AS respectively and the others human AS. The rate is calculated as the addition of all the spikes during a sound stimulus, divided by the length (time) of the stimulus and the amount of fibers.

Obtaining the phase locking

To characterize the phase locking, the synchronization index along the interest frequencies (from section 2.3) have been performed with 30 dB SPL over threshold at 10kHz and 1000 HSR fibers. The synchronization index (or vector strength (r)) is the normalized estimation of the neuron's tendency to fire at a particular phase in a stimulus cycle, and is formulated as:

$$r = \frac{\sqrt{[\sum_0^{K-1} R_k \cos 2\pi(k/K)]^2 + [\sum_0^{K-1} R_k \sin 2\pi(k/K)]^2}}{\sum_0^{K-1} R_k} \tag{2.6}$$

where $\mathbf{K}$ is the number of bins in the period histogram, and $\mathbf{R}_k$ is the magnitude of the $\boldsymbol{k}\mathbf{th}$ bin (Rees and Palmer, 1989).

How to get the modulation gain

To calculate the degree of modulation, 1000 HSR fibers have been used. The input signal consists of an amplitud modulated (AM) signal of 10 dB SPL over threshold at 10kHz, with a fc (carrier frequency) of 10kHz and a modulated frequency that varies from 100Hz to 2kHz (using values of 2.3). However, it is known that, for human, the modulation at low frequencies is even more important than at high frequencies. The nature of this low frequencies generate the need of a signal that confirms the duration of at least one period for a correct result. Therefore, the signal for low fm (modulation frequency), has the slope shown in fig. 2.5. Until T/4, when T is the period, the signal is a non-modulated toneburst, with a ramped start. The calculation of the MTF (modulation transfer function) takes just one period, from 3T/4 til 7T/4 (in the figure, the solid red arrow). The model of Wang et al. (2008) has not been tested at low frequencies because it computes the response for diferent CF and it has been impossible to process so long signals. The MTF (in dB) is calculated as the amount of percent modulation depth of the histogram divided by the modulation depth of the stimulus.

$$Modulation\ gain = 20 * Log_{10}\frac{200 * r}{m} \tag{2.7}$$

Where **r** is the vector strength as obtained in Eq. 2.6 and **m** is the modulation depth of the signal (in this thesis m=0.99) (Rees and Palmer, 1989).

How to obtain the tuning curves

The threshold along the frequency map (section 2.3) for a fixed CF (2kHz) has been obtained for 1000 HSR fibers. It has been computed as the sound level needed to obtain 10 spikes per second more than the spontaneous rate (Zhang et al., 2001; Zilany and Bruce, 2006). The signal used is a tone burst of the same frecuency as the CF and a variable sound level.

Obtaining the Q_{10dB} values

The Q_{10dB} is defined as the filter BF divided by the bandwidth at 10 dB over the threshold at BF (Miller et al., 1997).

$$Q_{10dB} = \frac{BF}{BW_{10\ dB}} \tag{2.8}$$

To get the Q_{10dB} along the frequency array (section 2.3), the tuning curves at these frequencies have had to be obtained. Therefore only 100 HSR fibers have been used for this test.

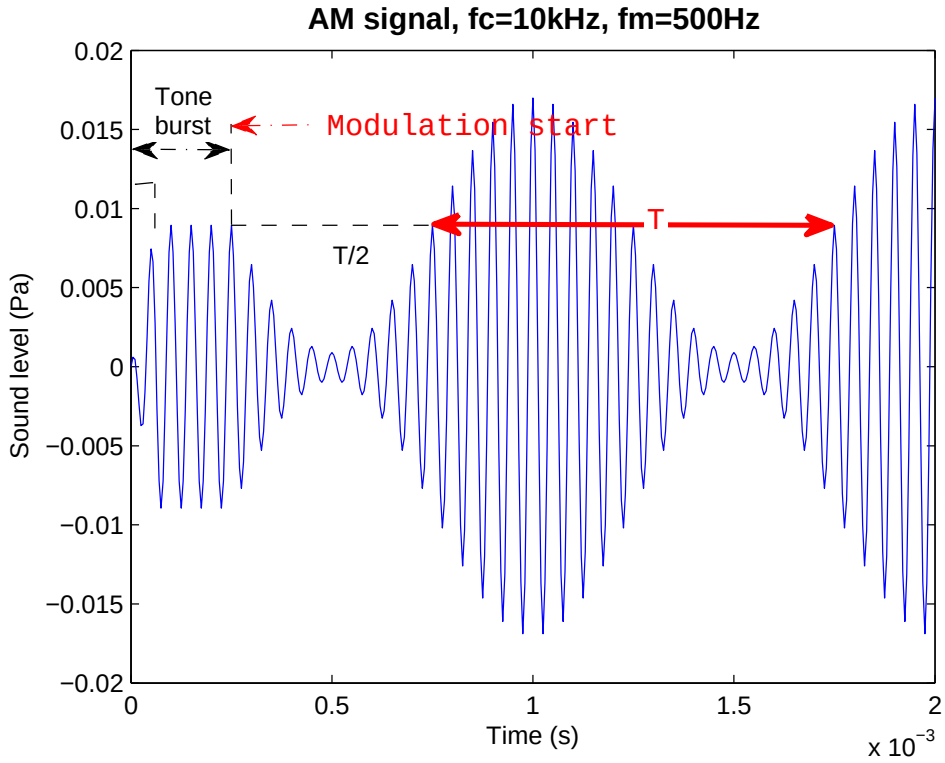

Figure 2.5: Example of an AM signal used for MTF. fc=10kHz, fm=500Hz, m=0.99. In T/4, when T is the period, the modulation start. The red solid arrow shows the amount of signal, one period, used to calculate the modulation gain

2.5 Performing the ASR

For the evaluation of the response for speech an ASR has been applied. To use the ASR, the spike signal from the models had to endure some modifications. Since the input data is notoriously redundant, it is transformed through feature extraction, a form of dimensionality reduction, into features vector. Thus, the input carries the linguistic information and supresses irrelevant acoustic information. To do that correctly, first a window of 25ms with a step of 10ms is used to segmentate the signal. Then it is applied a filter bank and a digital cosine transformation (DCT) to reduce the spectral resolution and decorrelate the features vector. From this DCT, only 12 components are used and the first and second order derivatives are added. That is the procedure of the front-end, one of the two parts of the ASR.

The other part of the ASR, the back-end, discriminates and classifies between classes through hidden Markow models. The speech recognizer is built with the Cambridge's HTK and uses multi-layer perceptrons. As the source used is the ISOLET database (Cole et al., 1990) and a version of ISOLET with noise and both contains the speak of the whole alphabet twice by 75 female and 75 male, the classes to be distinguished will be letters. The noisy ISOLET includes no noise and 20, 15, 10, 5 and 0 signal to noise ratio (SNR) (Holmberg et al., 2005). The noisy types comes from the RSG-10 collection (Steneken and Geurtsen, 1988). Both clean and noisy ISOLET were scaled by the same value to have physically meaningful amplitudes. Thus, the dynamic range of all the recordings was scaled but not the different sounds. The whole proccess is widely described by Holmberg (2007). 75 HSR fibers and 25 LSR fibers per frequency channel were selected to prepare the AN response which would be the input of the ASR system. The total of frequency channels were 100, using the total of human frequency map given in section 2.3. Thus, the total number of fibers across frequency is 10000 fibers. Taking into account that Holmberg (2007) pointed that increasing the amount of fibers over 1000 has a negligible effect, 10000 is more than it is necessary to get a good speech recognition, even for noisy stimulus.

Chapter 3

Results

This chapter discusses the result of testing the different models through a variety of analysis. The first section reviews the response of the stimulus in the presence of a tone burst as an input signal. Frequency and time analysis help us to understand the behavior of the models and give us the opportunity of compare them with experimental data. The second section examine the results of using the auditory models as an acoustic front-end processors for ASR system. Clean an noisy background have been used to evaluate the accuracy of the models in front of environment diverseness.

3.1 Simple tone stimulation

3.1.1 Temporal analysis with PSTH

Fig. 3.1 shows post stimulus time histogram for a frequency of 4.3kHz. With the PSTH, the difference in timing of the different model can be systematically investigated. Moreover, the rapid and short-term adaptation (two exponential decay functions) can be calculated. They where characterized by Westerman and Smith (1984) as two of the addend (the steady state constant was the third) of the adaptation function.

The PSTH of every model started with spontaneous activity for no signal. Then at 2 ms the signal starts and the onset-peak occurs at approximately 3ms, depending on models. The value of this peak changes considerately, from 600 sp/sec of the model of Wang et al. (2008) until more than 1500 sp/sec of the models of Carney's group. Besides, the model of Zilany et al. (2009) has a slim onset-peak (also valid for Heinz et al. (2001) model) and a notch after it, sometimes visible in AN fibers data (Ruggero and Semple, 1991). On the other hand, the models of Meddis' and Hemmert's groups have

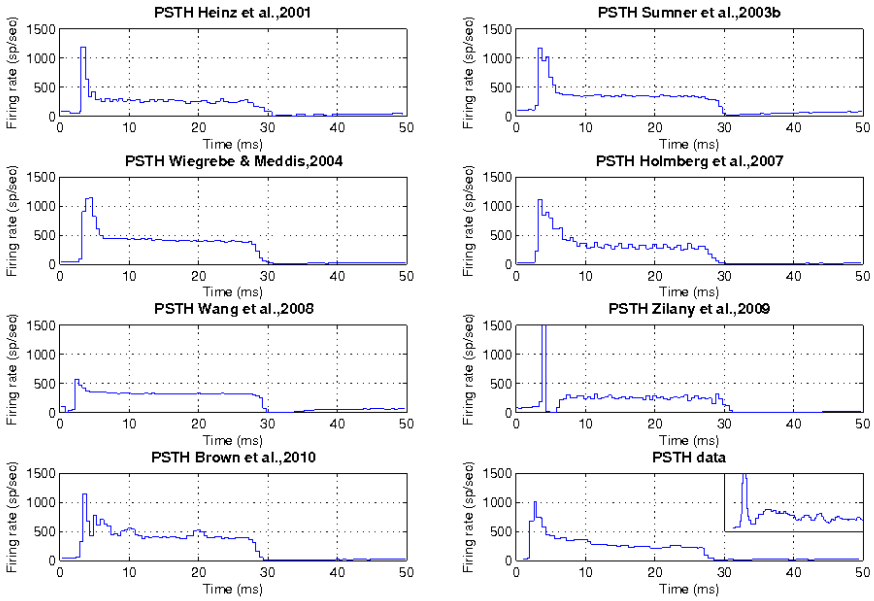

Figure 3.1: PSTH for 4.3kHz (blue line) and 10kHz (red line) signal input of 25 ms with 1 ms of rise and fall time and 30 dB SPL over threshold. The test was made for 1000 HSR fibers and a bin size of 0.5ms. The data (guinea-pig) is obtained from Muller and Robertson (1991). In the upper right corner of the data plot we find a detail of the notch after the onset peak (from Ruggero and Semple (1991)).

a wide peak, which width depend on intensity, that decreases exponentially (as explained above) to rise the sustained discharge rate. Though, the adaptation function given by the model of Holmberg et al. (2007) has a slope less sharp than the others. The steady state value lies, for almost all the models, around 300 sp/sec, although the model described by Zilany et al. (2009) has a lower value (near 200 sp/sec) and the model of Wiegrebe and Meddis (2004) has a higher one (almost 500). This last model has abnormal peaks at 5, 10 and 20 ms that could be a reaction to the delay of the efferent system that affects the temporal properties of the model. The signal lasts 25ms, and then the result has drops more or less pronounced depending on the model. The time of recovery before the model fires again the spontaneous activity is known as offset adaptation. This adaptation is a property of the AN-IHCS, limits some aspects of the temporal coding and produces a greater sensitivity to transient stimuli. In the model of Sumner et al. (2003b) and Holmberg et al. (2007) this recovery happens after less than a few miliseconds. This quick recovery does not fit with the physiological data that present a "dead time" indeed presented in Wang et al. (2008) model. Nevertheless, for the model of Zilany et al. (2009), this recuperation (after this "dead time") takes more time.

3.1.2 Rate intensity functions of the AN

As said in the introduction, the mammals have three different kind of AN fibers, high, medium and low spontaneous rate fibers. The HSR has activity in the absence of any stimulus, requires a small quantity of sound pressure level to fire above spontaneous activity and a small dynamic range. LSR on the contrary, has almost no spontaneous activity, a higher threshold and a slow rise once the threshold is exceeded. These properties can be observed in Fig. 3.2, where the figure shows us the quantity of spikes per second for the different models through the different sound level of input signal. It can be observed that, for the HSR fibers, the models have usually a threshold between 5 and 20dB SPL; for MSR fibers a threshold between 10 and 30dB SPL and for LSR fibers between 10 and 40dB SPL. Therefore, the results fit with the experimental data, the solid gray lines (Winter and Palmer, 1991).

Although the spontaneous rate of the HSR fibers should be high, the Brown et al. (2010) model gives a result of only 20 spikes/sec. Nevertheles, it is important to emphasize that, some parameters from the models of Meddis group (Ca^{2+}_{thr}, M_{max} and G^{max}_{Ca}) can be changed to fit any guinea-pig rate intensity data as explained in Sumner et al. (2002).

Other characteristic observable is the dynamic range. While some models as the ones described in Sumner et al. (2003b), Zilany et al. (2009) and Brown

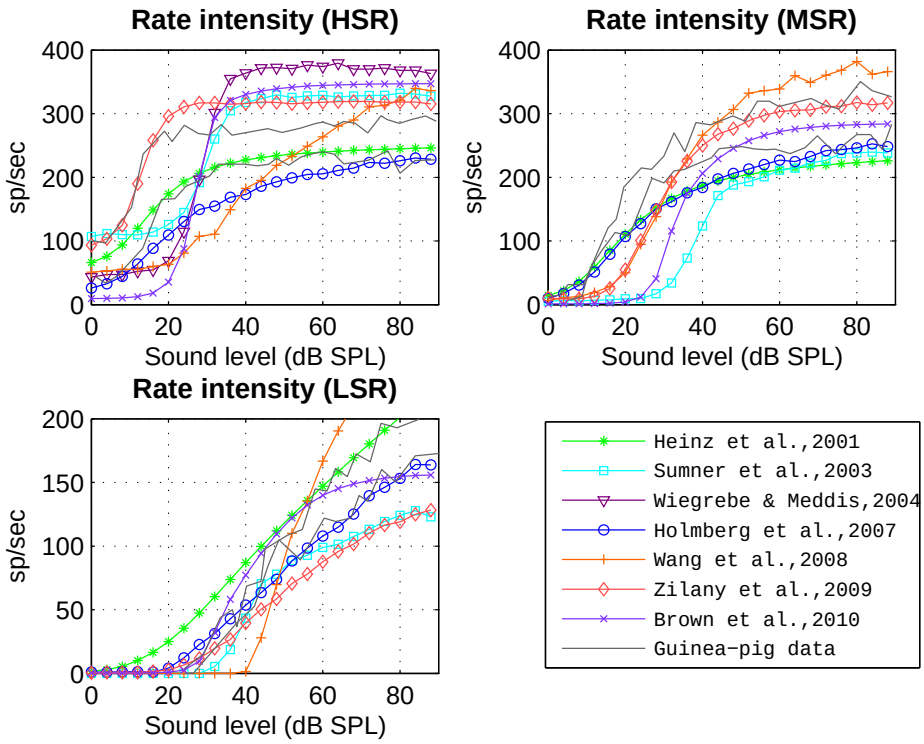

Figure 3.2: Rate intensity response for the seven models tested. The three different kind of fibers are tested separately. 250 fibers of each type have been used with a CF of 14kHz for the model of Sumner et al. (2003b) and Zilany et al. (2009), and a CF of 4.3kHz for the models presented in Heinz et al. (2001), Wiegrebe and Meddis (2004),Holmberg et al. (2007), Wang et al. (2008) and Brown et al. (2010). The guinea-pig data, in dashed lines, comes from Winter and Palmer (1991).

et al. (2010) have really pronounced HSR slopes once the threshold has been overstepped (15 dB of dynamic range), the rest of the models tested have a smoother one. In particular the models of Hemmert group have more than 50 dB of dynamic range. That does not happen in MSR fibers, at least it is not as remarkable as in HSR. In MSR fibers almost all the models, but the ones presented in Holmberg et al. (2007) and Heinz et al. (2001), have a dynamic range of 30dB approximately. For LSR fibers, the difference in dynamic range between models is even less different. However, the model of Wang et al. (2008) that was showing a really smooth response in HSR and MSR fibers, shows here the smallest range.

Sumner et al. (2003b); Wiegrebe and Meddis (2004); Zilany et al. (2009); Brown et al. (2010) models for HSR fibers conform better with the experimental data, that present a small dynamic range and a threshold around 5 dB SPL. However, we must not forget that the models presented in Heinz et al. (2001); Holmberg et al. (2007); Wang et al. (2008) simulate human ear and the data to compare are guinea-pig data. On the contrary, for the MSR and LSR fibers are not as clear as for HSR. The models, mostly, fit the experimental data of guinea-pig, though the model of Wang et al. (2008) has too high dynamic range in LSR fibers and a high rate in MSR.

3.1.3 Synchronization index of AN fibers as a function of CF

The indicant of the phase locking, an esencial temporal property, is the synchronization index. It points if the AN response conserves or not the phase and time structure of the input stimulus. That can be observed in Fig. 3.3, where the different behaviour from two different animals, cat and guinea-pig, and the results of the models are shown. At low frequencies the data present an index near to 1 (the spikes are synchronizated with the stimulus), while at higher frequencies this synchronization index falls down to almost 0 (spikes are generated randomly). While the synchronization index in the guinea pig drops from 1kHz, it drops with a higher frequency for cats. This low pass behavior of the data is presented (to greater or lesser extent) in the models. The Sumner et al. (2003b) model that simulates the guinea-pig ear, fits the data from Palmer and Russel (1986). The results from this model of Meddis group stays close to the average of the data, however it presents a non common response between 100Hz and 300Hz. This hill turns into a valley if a lower sound pressure level is used due to the ME filter. The gain at this frequencies is really low and therefore a lower sound level would be under the threshold. Fitting the data from cats (Johnson, 1980), the model

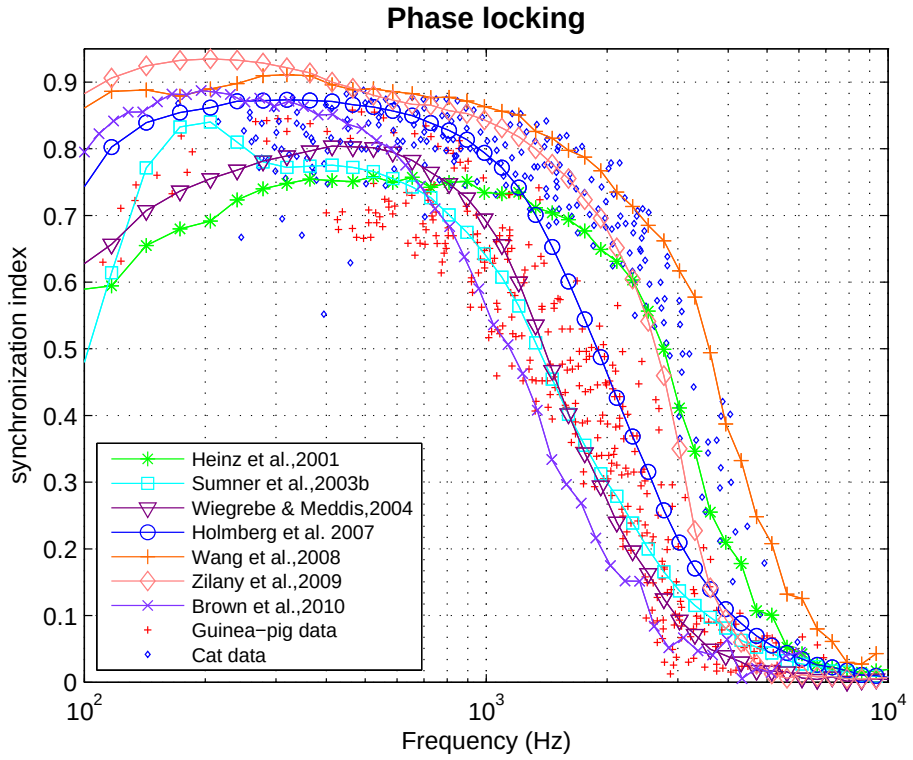

Figure 3.3: Sinchronization index along frequency for 1000 HSR fibers and a signal of 30 dB SPL over threshold. The black crosses are guinea-pig data from Palmer and Russel (1986), the red diamonds are cat data from Johnson (1980)

from Zilany has a good phase-locking response from 400Hz, although it has a pronounced fall from 1kHz. This shape is mainly the result of the PLA used that increases the synchronization. It also has a hill up to 400Hz due to the model is not too realistic at low frequencies.

Without data about human phase-locking, only a comparison between another animal data can be made. The best result obtained corresponds to Wang et al. (2008) model, its drop has the highest frequency and besides, it also has a good synchronization index at low frequencies. It improves the result of the previous model of Hemmert group, Holmberg et al. (2007), whose result stays between the cat and guinea-pig data. The models described by Sumner et al. (2003b) and by Wiegrebe and Meddis (2004) have a similar synchronization index as expected, since the both models are quite similar. With reference to the model of Brown et al. (2010), it has the drop at the lowest frequencies and a good phase locking at low frequencies. Completely the opposite of Heinz et al. (2001) model, which has a low synchronization index at low frequencies (less than 0.8) but a high drop frequency.

In general, the models fit the experimental data (cat or guinea-pig) with varying degrees of success but in any case comply with the physiological results.

3.1.4 Response to AM tones. The MTF from 0.1Hz to 2kHz

The MTF offers the synchronization of the AM signals to the fm (Fig. 3.4). As pointed in Greenwood and Joris (1996), mechanical and temporal filtering limit the cut-off frequency above which AN spikes response is not modulated by fm anymore. The mechanical band pass filter is generated by the local basilar partition motion driving the IHC and the temporal filtering reside between mechanical motion and the AN spikes and limits the AN synchronization to temporal variation in the IHC input.

Using the CF as fc of the input signal, the sideband components of the AM is removed by the mechanical filter as fm increases. This effect generates in turn a reduction in the envelope amplitude variation that influences the modulation gain cut-off frequency. Therefore a function of CF, sets the bandwidth of this filter. It has been related that high Q_{10dB} produce lower cut-off frequencies as happens to the models of Holmberg et al. (2007) and Wiegrebe and Meddis (2004). But the mechanical filter is not the only one that affects the shape of the MTF. Fixing the temporal filter with a high cut-off frequency results in a high cut-off frequency of the modulation gain. That is the solution of Zilany et al. (2009) to replicate accurately physiological

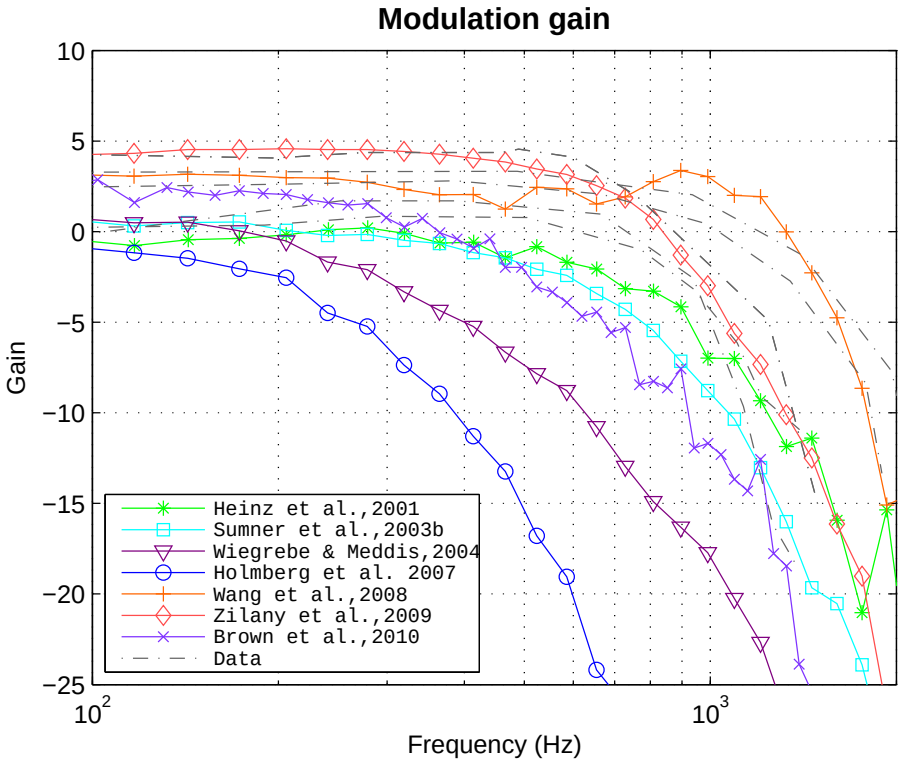

Figure 3.4: MTF for 1000 HSR fibers for the different models. The MTF is calculated as the amount of percent modulation depth of the histogram (200*r) divided by the modulation depth of the stimulus. The dashed lines are data from Joris and Yin (1992)

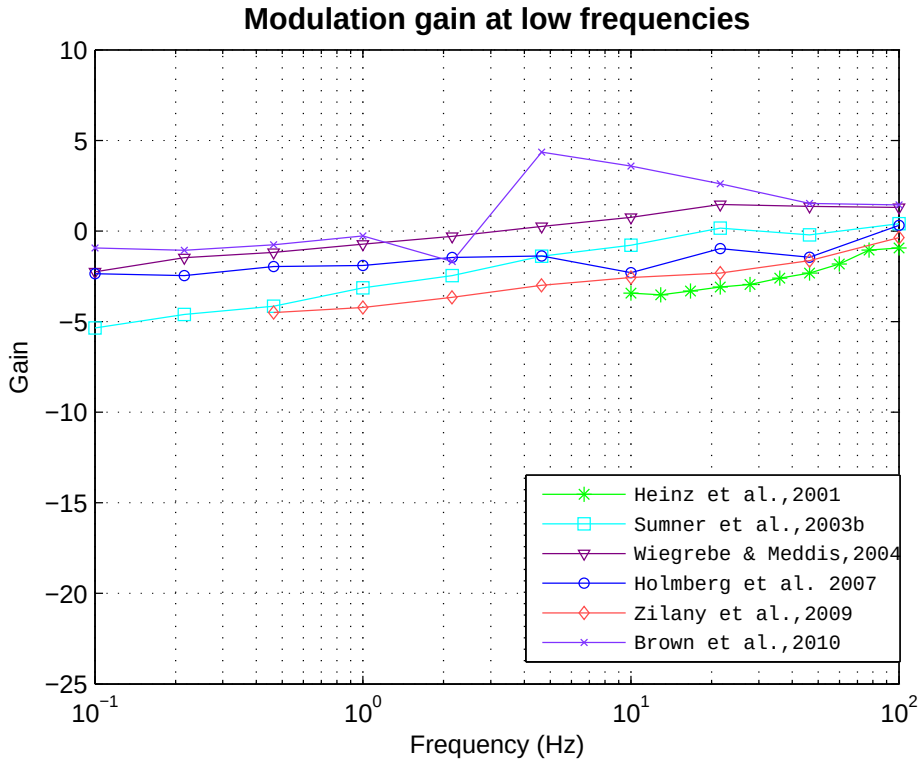

Figure 3.5: MTF at low frequencies for 1000 HSR. The amount of AM signal used to calculate is fixed to 1 period

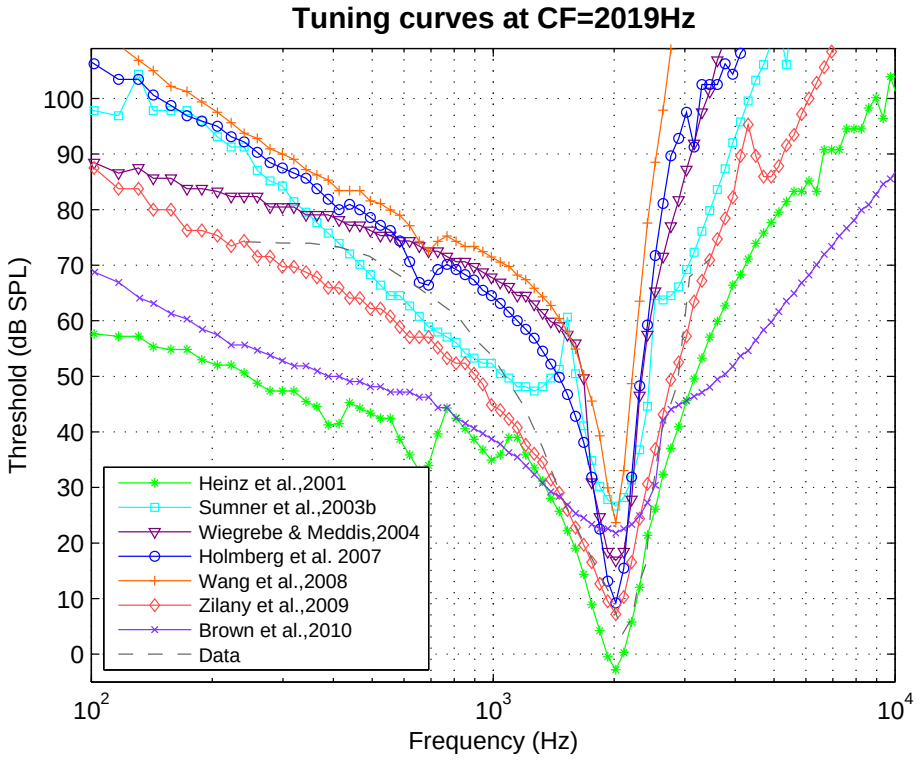

Figure 3.6: Minimum threshold to obtain 10 spikes/sec more than spontaneous rate across a range of frequency for a fixed CF=2019Hz and 1000 HSR. The data comes for cat from Liberman and Kiang (1977)

MTF. On the other hand, the Wang et al. (2008) model that includes an offset adaptation stage in their model get a higher cut-off frequency due to the OA properties of the synapse that increase the synchronization. These two models are the only ones that comply precisely with the data of Joris and Yin (1992), despite their differences, due to the differences in the same data. Besides, for speech recognition it is also important to look at very low frequencies (under 100 Hz). In fig. 3.5 the MTF of the models at very low frequencies is showed, despite the nonexistence of data at these frequencies. Therefore, we can only assume the result of the models and compare between them. The modulation gain slightly increase (3 dB max) from 0.1Hz to 100 Hz and are quite uniform for almost all the models. However, the model of Brown et al. (2010), with its increase from 2Hz to 40Hz, is the only exception.

3.1.5 Frequency threshold tuning curves

The minimum threshold at which the nerve fiber will respond depends on the frequency of the input stimulus. It is easily visible on plot Fig. 3.6 the asymmetry of the curve. How the threshold stays over [60–100] dB SPL (depending on models) for frequencies under CF, how it falls to almost 0 at CF (depending on the ME filter and IHC gain) and increases quickly for higher frequencies. It also helps to indicate the selectivity and sensitivity of the model with a simple look.

The bandwidth of the tuning curve is a good indicator of how the frequencies that are not the CF will be attenuated. Due to this reason, to recognize the model with the best selectivity could be a good start to point out the best model for ASR. However, into the previous process of the ASR, specifically in the DCT where only 12 components are used, this selectivity will be lost. For this reason, it is not relevant that the model of Brown et al. (2010), particularly designed for ASR shows a wide tuning curve. The models of Heinz et al. (2001), Zilany et al. (2009) and Sumner et al. (2003b) present a narrower result, although the narrowest bandwidth goes to the model of Wiegrebe and Meddis (2004), Holmberg et al. (2007) and Wang et al. (2008).

It is important to mention the singularities of the different models presented in this plot, such as the peak of Sumner et al. (2003b) model at 1.5kHz and the peak of Zilany et al. (2009) model at 4.3kHz or the valley of the model of Holmberg et al. (2007) and the one of the model described in Wang et al. (2008) at 750Hz. Some of these anomalies are due to the mode of development and programming. However, this is not the case of the result of the model of Zilany et al. (2009), which anomaly correspond to the ME notch above 4kHz. As expected, the results of this model are the ones that fit better with the cat data of Liberman and Kiang (1977) because is the only model that mimic cat APS. On the other hand, the models that try to simulate human ear and have achieved a narrow tuning curve (specifically the ones described in Wiegrebe and Meddis (2004); Holmberg et al. (2007); Wang et al. (2008)), fit the psychophysical tuning curve data of Carney and Nelson (1983).

3.1.6 Q_{10dB}, a selectivity measurement

The Fig. 3.7 represents the Q_{10dB}, a measure of the sharpness of tuning, throughout frequency. With it, it is even easier to point the selectivity of the model because it represents the BF divided by the bandwidth of the tuning curve at 10 dB over the minimum threshold. As noted in the section 3.1.5, this selectivity would become unprofitable after the DCT stage. Although

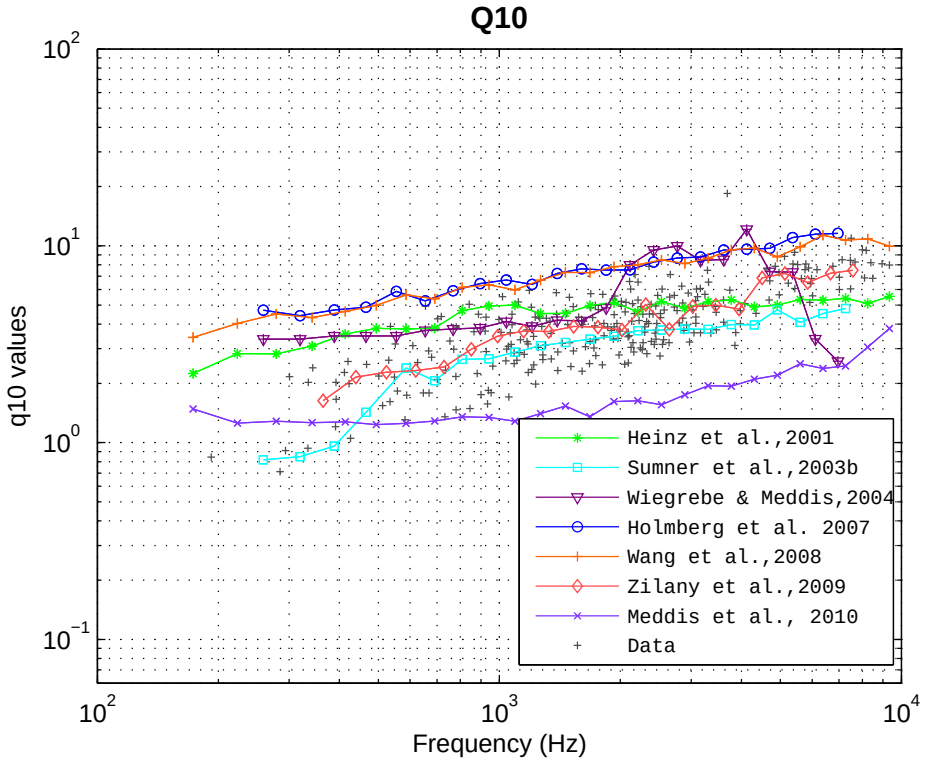

Figure 3.7: Q_{10dB} along frequency for 100 HSR fibers. Grey crosses are cat data from Miller et al. (1997)

almost all the models present unsurprising results (previous study of the Fig. 3.6) with a high Q_{10dB} for Wang et al. (2008) and Holmberg et al. (2007) models, medium for Sumner et al. (2003b), Zilany et al. (2009) and Heinz et al. (2001) models and low for Brown et al. (2010) model; the Wiegrebe and Meddis (2004) model shows a low Q_{10dB} up to 2kHz and above 6kHz, that represent the maximum CF of the model, and a higher Q_{10dB} between these both frequencies. The cat model of Carney group is the one that fits more properly the data of Miller et al. (1997). On the other hand, the human models of Holmberg et al. (2007); Wang et al. (2008), and even the model of Wiegrebe and Meddis (2004) for certain frequencies, have a Q_{10dB} that exceed the higher value of cat data. Nevertheless the psychophysical data of Carney and Nelson (1983) have characterized a narrow tuning curve that in turn means a higher Q_{10dB}. Therefore, it is easy to point out that they will fit the human data.

3.2 Automatic speech recognition

Once the models were surveyed for simple tone stimulation, a study about the results for complex stimuli was conducted. Due to the computational complexity of this part, I only use the two models that can be more relevant, the ones described by Wang et al. (2008) and Zilany et al. (2009).[1] Here we will see how the background noise and the sound level of the input signal affect the recognition percent.

3.2.1 Effects of noise level

In this part of the thesis, speech recognition accuracy was obtained as a function of SNR. 20,15,10,5 and 0 signal to noise ratio and a clean signal were used. As seen in fig. 3.8 the result of both models is a monotone decreasing curve. The two models present a similar response for the best result (40dB SPL for the model of Zilany et al. (2009) and 70 for the model of Wang et al. (2008)) amount a variety of input signals. Though Zilany et al. (2009) has achieved usually a better recognition, the higher difference between the two models has been less than 5 percent (for a 0 SNR).

[1]The model of Brown et al. (2010) is indeed the third one that should be tested. However, the results of this model for ASR have been already conducted in the same publication and the improvement in recognition also pointed out.

3.2.2 Effect of sound level

To know how well the models recognition will be in the presence of different sound levels, it is important that the following experiment is performed. In fig. 3.9 speech recognition accuracy has been obtained for different configuration of the models in which the sound level pressure of the speech signal was set to 10,40 and 60 dB over the threshold of this models at 10kHz. For each sound level, speech recognition accuracy has been evaluated in a range of noise conditions. The results of the model described by Wang et al. (2008) show a high difference between the line at 10dB and the 40 and 60 dB for clean conditions. However, this difference decreases when the noise raises and since 10 dB of SNR the results deviate about 5 percent. That does not happen for the results of the model of Zilany et al. (2009), that maintain more or less the same amount of difference between sound levels for each signal to noise ratio. For signals of low sound level, the results of this model have proven to be better with clean conditions or low noise background. On the other hand, when the noise increases to the same level of the signal, the model of Wang et al. (2008) has got the best result. However, that is a particular case. In the fig. 3.10 we can see that the model of Zilany et al. (2009) in noisy conditions (the average of the result for 20,25,10,5 and 0 SNR) has a better result for signals with a low sound level up to 60 dB SPL. After that, the model described by Wang et al. (2008) gets a higher percentage. This pattern is repeated in the clean condition, but at 70 dB SPL. Once again, the model of Carney group gets a better result for low sound level while the model of Hemmert group has a better result for high sound levels. One might think that this result is the fruit of the different thresholds in the models that ensues in a displacement in the curves. But after normalizing the two curves is still observed this behavior, model from Zilany et al. (2009) performs better for low intensity sounds and model from Wang et al. (2008) for high intensity sounds.

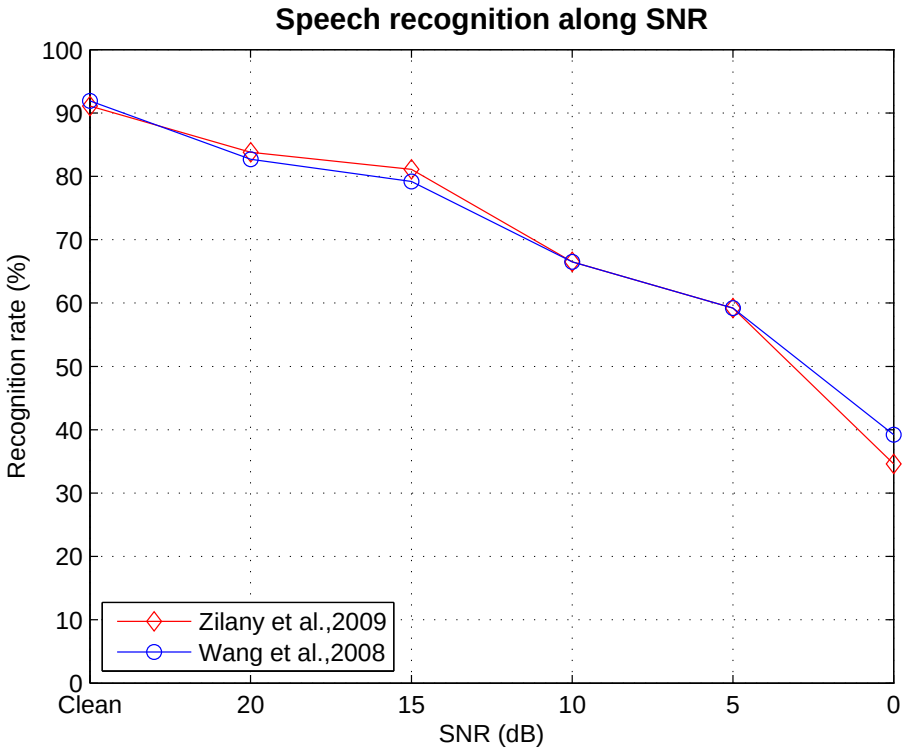

Figure 3.8: Speech recognition for different noise background. The sound level used was the one that got the best results (40dB SPL for the model of Zilany et al. (2009) and 70 for the model of Wang et al. (2008)).10000 fibers, 75 HSR and 25LSR per frequency channel were used

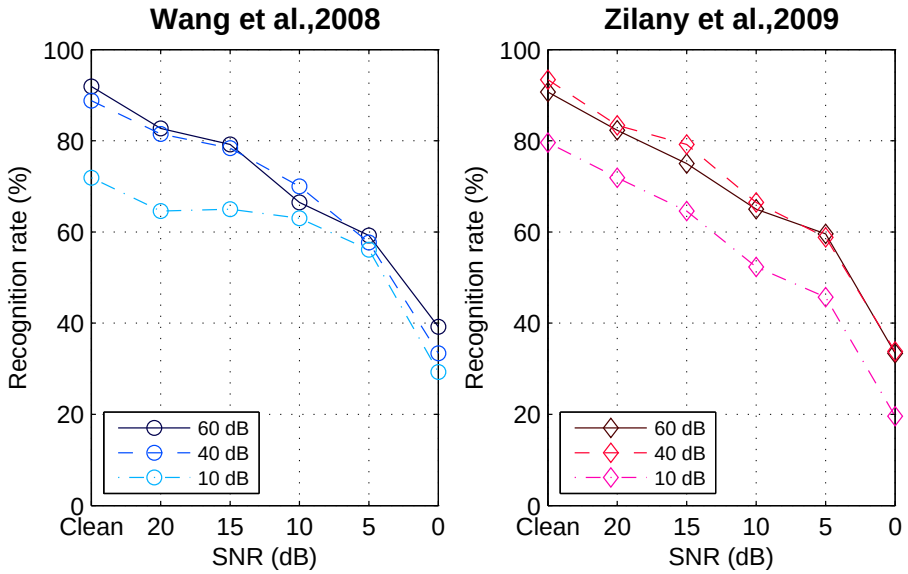

Figure 3.9: Percent of recognition for different sound level over threshold at 10kHz and a variety of signal to noise ratio. A total of 10000 fibers (3/4 of HSR and 1/4 of LSR) were used. Left: model of Wang et al. (2008) (with a threshold of 20dB SPL). Right: model of Zilany et al. (2009) (with a threshold of 10dB SPL)

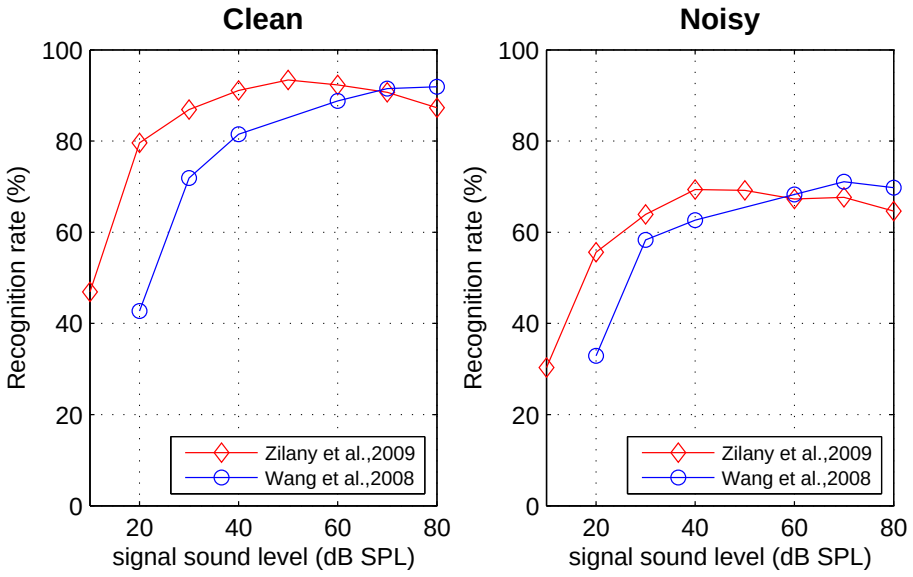

Figure 3.10: Clean (left) and noisy (right) conditions for the models of Wang et al. (2008) and Zilany et al. (2009) throughout sound pressure level and 100 fibers (HSR and LSR) per channel. The noisy plot represents the mean of the results under a variety of noise conditions (20,15,10,5 and 0 SNR)

Chapter 4

Discussion

This study presented a comparison between a variety of auditory periphery models from three different research groups. The main motivation was to compare the models against experimental data first and then to use them as an ASR backend to discover how the different properties and implementations of those models influence ASR results. Thus, we can conclude which properties of auditory periphery system are relevant for speech recognition. To achieve this goal, we have analyzed the models with simple and complex stimulus. With simple signals, we have explored seven models with different evaluation, we have analyzed their properties and then we have compared the results. The timing properties with the PSTH, the dynamic range and threshold with rate intensity plots, the selectivity with the tuning curves and Q_{10dB} values and the phase and time structure in presence of tone burst or AM signals with synchronization index, were the attribute examined. The complex stimulus were speech signals. The AN fibers output from the models in presence of speech were used as front end of an ASR system. This is not exactly new, other publications as Holmberg et al. (2007); Wang et al. (2008); Brown et al. (2010), referring only to those studied here, have done it before. Holmberg et al. (2007) tried to test speech encoding strategies while Wang et al. (2008) and Brown et al. (2010) tried to demonstrate the worth of using an OA function or efferent activity attenuation through ASR respectively. The work presented here can not be directly compared with the work of Brown et al. (2010) because they used a different database of speech signals. On contrary, the work of Holmberg et al. (2007); Wang et al. (2008), that use this database, have been a reference. Although the number of models tested with the simple signals were seven, only two models were used in this part of the thesis because the large number of speech signals in the ISOLET database entailed a high consume of time.

Studying the results

Looking in the post stimulus time histogram, we found the first differences among models. On the one hand, the result after evaluating the model of Brown et al. (2010) showed some uncommon peaks probably as a result of the efferent delay. On the other hand, the onset peak of Wang et al. (2008) is not as high as the rest of the models even for high sound levels (30 dB SPL over threshold). This model and the one described by Zilany et al. (2009), take care over the offset adaptation, they showed a "dead time" after the stimulus like in experimental data.

In the rate intensity functions of the AN fibers we discovered a high and unrealistic dynamic range of the model described by Wang et al. (2008) due to the offset adaptation. There is also a high threshold of Brown et al. (2010), the shift in the rate intensity response is a consequence of the efferent activity modeled as an attenuation. However, all the models from Meddis group have showed a higher threshold in HSR fibers. Generally, the phenomenological models (as Heinz et al. (2001); Zilany et al. (2009)) have achieved the results that better fits with the data.

The different animal data used in synchronization index was a helpful tool to study the structure of AN fibers response. This study has pointed out the outperforming of models that includes some kind of offset adaptation. I mean, the models described by Wang et al. (2008); Zilany et al. (2009). However, the second model has increased its synchronization index in an fanciful way for low frequencies.

Something similar occurs in the evaluation of MTF. Looking at fig. 3.4 is easy to see the ameliorated result of models of Wang et al. (2008) and Zilany et al. (2009) although this last model have a smoother result. On contrary, the result at very low frequency have not pointed out any difference among models but the one from Brown et al. (2010). Since there is no experimental data at these low frequencies, it is difficult to select the correct one.

The tuning curves and the Q_{10dB} values highlight the bad selectivity of Brown et al. (2010) model. Since the model was intended to work with an ASR that uses only 14 components of the DCT, is probable that the authors disdained the selectivity in their model. On contrary, the models of Holmberg et al. (2007) and Wang et al. (2008) that also uses an ASR with 12 components, have a very narrow tuning curve and therefore high Q_{10dB} values that would fit the human psychophysical data. It is important to mention that, the model described by Wiegrebe and Meddis (2004) have a huge difference in the Q_{10dB} values along frequency without plausible cause.

Concerning the ASR, only the results of Wang et al. (2008) and Zilany et al. (2009) models were studied here. Both models simulate the offset

adaptation, the first with the shift described in Zhang and Carney (2005) and the second as a consequence of its power law adaptation at the synapse. Perhaps for this reason, the expected difference between the two models recognition (for the sound level with better results) did not exist at the end. This would mean that the other implications of using a PLA do not affect the ASR results for a convenient sound level or the effect is insignificant. Not so insignificant when you consider low intensities, where the high dynamic range of the model described by Wang et al. (2008) is not a realistic result. It could have been the cause of the lower recognition and thus, the model of Zilany et al. (2009) and their PLA had a better result. With a higher dynamic range, the amount of spikes produced at low intensities even above threshold is much lower. However, it is also possible that the different way of the intrinsic implementation of both models would be the responsible of the this difference. By the same token, neither be tuned for different animals (the model of Zilany et al. (2009) mimics cats and the model of Wang et al. (2008) mimics human ear), nor to have a better or worse selectivity has supposed a difference for ASR.

In any case, both models have achieved a better result than models without offset adaptation performed, as the model of Holmberg et al. (2007). Thus, we can conclude that offset adaptation is crucial for speech recognition in clean and noise environments and the implementation of this offset adaptation, shift in AN response (Wang et al., 2008) or PLA (Zilany et al., 2009) does not affect the result. Besides, taking into account all the results from simple tone stimulation we also conclude that the model that fits better with the data is the one presented by Zilany et al. (2009). The fail in the synchronization index at low frequencies and high intensities is only a very specific situation that does not affect that this model has the best result in all the test done.

The better improvements

Among all the improvements performed by the different models along the years, there are some of them that, to this day, are essential. The first indispensable improvement that has been evident is an amelioration in the offset adaptation. Not only the offset adaptation is thought the responsible of psychophysical forward masking (Harris and Dallos, 1979), but the models which consider it (as the model of Wang et al. (2008)) increases the synchronization index and the coding of amplitude modulated signals. Those improvements with the offset adaptation were expected by Holmberg et al. (2007) and Zhang and Carney (2005) among others. In the same way we can find the power law adaptation of Zilany et al. (2009) As pointed in Zhang and

Carney (2005), adaptation occurs (or it is believed at least) at the level of the transmitter release process in the AN-IHCS that is exactly the component changed in this model. With the power law adaptation, the offset adaptation as well as the adaptation to increments and decrements in a ongoing stimulus, can be simulated. This improvements about the offset adaptation would be the responsible of the good synchronization's result of the model of Zilany et al. (2009) like the model of Wang et al. (2008). Both models have achieved the best synchronization index (Fig. 3.3) and the best modulation gain (Fig. 3.3) among the other models evaluated.

Another crucial improvement is the use of ME filter and/or eardrum transfer function. On the one hand, the results of the model of Heinz et al. (2001) (without ME filter), have shown that having or not a ME filter does not affect the results, except for the threshold at which the fibers will fire over the spontaneous activity. On the other hand, this model has only been studied with simple signals. ME has demonstrated its significance in modeling responses to wide-band stimuli, where the difference in gain along frequency would be relevant. More than an effect of eradicating malicious noise is intended to approach the model to the physiological process, boosting or reducing the amplitude of the signal depending on the frequency, as does the ear. Furthermore, the computational complexity and time needed to mimic the effect of the middle and outer ear is negligible.

By the same token, the models should be able to use high, medium and low spontaneous fibers. The implication of developing different kind of fibers might seem not as trivial as the ME filter. The truth is that mostly, it only involves a change in the parameters used in the part of the model corresponding to AN-IHCS. The main importance of the use of LSR is its role in encoding high sound levels. As seen in fig. 3.2 at this high levels the other fibers are saturated. This coincides with the results obtained by Winter and Palmer (1991). Moreover, LSR seems to be important for vowel coding (Holmberg, 2007). In this research work the author noticed that LSR features surpass HSR ones at normal conversation sound levels in a experiment with only vowels.

The last improvement that has been compared is the use of efferent activity by attenuating the BM response. For simple signals in a clean environment this improvement has nothing to contribute because its strength is the antimasking effect in noisy environments (Ferry and Meddis, 2007). Thus, the phase locking and the MTF of the Brown et al. (2010) model did not show any improvement. The only appreciable change for this simple stimuli is the shift in the rate intensity response (fig. 3.2), due to the attenuation that the efferent activity implies. Moreover, the result of the tuning curve and Q_{10dB} show an unfortunate surprise: a bad selectivity. However, the selectivity of

the reproduction of the model presented by Brown et al. (2010) has nothing to do with the efferent system since the model has the same output when the use of efferent activity is turned off. It seems pretty clear that, although the implementation of the efferent activity is quite easy (Ferry and Meddis, 2007), the use of efferent activity is not needed if the model is not intended to work in a noisy background, at least if this efferent activity is modeled as a simple attenuation.

In conclusion, the future models that would like to obtain an accurate AN response in presence of speech signals would have to use a ME filter and/or eardrum transfer function, different fiber types (at least HSR and LSR), a correct offset adaptation and an efferent activity feedback if it is intended to work in noisy environments.

Future work

This study has not considered an important effect: "the forward masking". The forward masking occurs when a signal arrives to the auditory system and it has its responsiveness reduced as a consequence of a previous acoustic stimulation. The OA could be responsible of this phenomenon as pointed above. The study of forward masking would be important in future comparisons.

The use, implementation and posterior comparison of the model described by Brown et al. (2010) has not got as good results as expected. The model has an extensive amount of parameters and is relatively new. Due to the variety of parameters files offered by the author, different tests to select the correct one were made. Then the parameters file which fits better with experimental data was chosen. However, although the model fits the experimental data, it has not achieved good results compared with the other models studied. In particular, in the tuning curves and the in the Q_{10dB} tests, it has been the worst one simulated. The model was intended to be a tool for evaluate the use of efferent activity, comparing the results of the model with and without it in presence of noise. It could be that the other aspects of the model have been despised.

It might be thought that two models are really very few to compare. At least three models, from the three research groups studied would be required. After all, no model of Meddis group was studied with speech signals in this thesis. However, the analysis with simple signals gives us a good basis of understanding above all the models and the last model from Meddis group had already been assessed with the use of ASR. Thereby, it can be concluded that, the comparison of these two models and the results published by Brown et al. (2010) are more than enough to accomplish the aim of this thesis. However, for a quantitatively comparison, the study of this model should be

made through an ASR with the same source used with the other models, I mean with the ISOLET data base.

In some cases there are not human data, at least not physiological ones. The psychophysical data not only represent the results that we intend to see but also the underlying superior process that could conduct to an erroneous clarifications.

The lack on time to evaluate of the models has been the main limitation. The process of obtaining the speech recognition along 7800 audio files was a high time consumer. Therefore only two models could be studied with ASR. In future studies, all the models studied for simple signals should be evaluated also with ASR to compare the results properly.

Conclusions

After the project is completed and all models have been studied among all the tests proposed, we are able to select the better improvement from the results of tests.

- The model of Zilany et al.(2009)is the one that fits better with the data. Although it has some inadequate temporal response at low frequencies and high sound level, it is only a very specific situation.
- Offset adaptation has demonstrated to be essential, both as a shift in the AN fiber response function and as being a result of the inclusion of power law functions in the synapsis.
- Selectivity and animal tuning are not relevant for speech recognition. At least if the ASR system uses only a part of DCT components.
- Efferent activity is useless in a clean environment, although the results of Brown et al. (2010) have demonstrated that is an important improvement for noisy background.

.

List of Figures

List of Tables

Bibliography

Brown, G., Ferry, R., and Meddis, R. (2010). A computer model of auditory efferent suppression: Implications for the recognition of speech in noise. *Journal of Acoustical Society of America*, 127(2):943–954.

Bruce, I., Sachs, M., and Young, E. (2003). An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. *Journal of Acoustical Society of America*, 113(1):369–388.

Carney, A. and Nelson, D. (1983). An analysis of psychophysical tuning curves in normal and pathological ears. *Journal of Acoustical Society of America*, 73(1):268–278.

Carney, L. (1993). A model for the responses of low-frequency auditory-nerve fibers in cat. *Journal of Acoustical Society of America*, 93:401–417.

Chittka, L. and Brockmann, A. (2005). Perception space. the final frontier. *Public library of science. Biology*, 3(4):564–568.

Cole, R., Muthusamy, Y., and Fanty, M. (1990). The isolet spoken letter database. Cs/e 90–004, Oregon graduate institute.

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

Ferry, R. and Meddis, R. (2007). A computer model of medial efferent suppression in the mammalian auditory system. *Journal of Acoustical Society of America*, 122(6):3519–3526.

Fettiplace, R. and Hackney, C. (2006). The sensory and motor roles of auditory hair cells. *Nature Reviews Neuroscience*, 7:19–29.

Greenwood, D. and Joris, P. (1996). Mechanical and temporal filtering as codeterminants of the response by cat primary fibers to the amplitude-modulated signals. *Journal of Acoustical Society of America*, 99:1029–1039.

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

Heinz, M., Zhang, X., Bruce, I., and Carney, L. (2001). Auditory nerve model for predicting performance limits of normal and impaired listeners. *Acoustics research letters online. Acoustical Society of America*, pages 91–96.

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

Holmberg, M., Gelbart, D., and Hemmert, W. (2005). Automatic speech recognition with neuronal spike trains. In *Proc. Interspeech'05*, pages 1253–1256.

Holmberg, M., Gelbart, D., and Hemmert, W. (2007). Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition. *Speech Comunication*, 49(12):917–932.

Johnson, D. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *Journal of Acoustical Society of America*, 68:1115–1122.

Joris, P. and Yin, T. (1992). Responses to amplitud-modulated tones in the auditory nerve of the cat. *Journal of Acoustical Society of America*, 91(1):215–232.

Kawase, T. and Liberman, C. (1993). Antimasking effects of the olivocochlear reflex. i. enhancement of compound action potentials to masked tones. *Journal of Neurophysiology*, 70:2519–2532.

Kiang, N. (1990). Curious oddments of auditory-nerve studies. *Hear Research*, 49:1–16.

Liberman, M. (1988). Response properties of cochlear efferent neurons: monaural vs. binaural stimulation and the effects of noise. *Journal of Neurophysiology*, 60:1779–1798.

Liberman, M. and Kiang, N. (1977). Tuning curves of auditory-nerve fibers. *Journal of Acoustical Society of America*, 61.

Lopez-Poveda, E. and Eustaquio-Martin, A. (2006). A biophysical model of the inner hair cell: The contribution of potassium current to peripheral auditory compression. *Journal of the Association for Research in Otolaryngology*, 7:218–235.

Lopez-Poveda, E. and Meddis, R. (2001). A human nonlinear cochlear filterbank. *Journal of Acoustical Society of America*, 110(6):3107–3118.

Meddis, R. (2006). Auditory-nerve first-spike latency and auditory absolute threshold: A computer model. *Journal of Acoustical Society of America*, 119:406–417.

Meddis, R. and O'Mard, L. (2005). A computer model of the auditory-nerve response to forward-masking stimuli. *Journal of Acoustical Society of America*, 117(6):3787–3798.

Meddis, R., O'Mard, L., and Lopez-Poveda, E. (2001). A computational algorithm for computing nonlinear auditory frequency selectivity. *Journal of Acoustical Society of America*, 109(6):2852–2861.

Miller, R., Schilling, J., Franck, K., and Young, E. (1997). Effects of acoustic trauma on the representation of the vowel /ε/ in cat auditory nerve fibers. *Journal of Acoustical Society of America*, 101(6):3602–3616.

Muller, M. and Robertson, D. (1991). Relationship between tone burst discharge pattern and spontaneous firing rate of auditory nerve fibers in the guinea pig. *Hearing Research*, 57:63–70.

Palmer, A. and Russel, I. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner-hair-cells. *Journal of Acoustical Society of America*, 24:1–15.

Rees, A. and Palmer, A. (1989). Neuronal responses to amplitude-modulated and pure-tone stimuli in the guinea pig inferior colliculus, and their modification by broadband noise. *Journal of Acoustical Society of America*, 85(5):1978–1994.

Ruggero, M. and Semple, M. (1991). Acoustics, physiological. *Encyclopedia of Applied Physics*, 1.

Seneff, S. (1985). *Pitch and spectral analysis of speech based on an auditory synchrony model*. PhD thesis, Massachusetts Institute of Technology.

Slaney, M. (1998). Auditory toolbox. Technical report, Interval Research Corporation.

Steneken, H. and Geurtsen, F. (1988). Description of the rsg-10 noise database. Technical report, TNO Institute of Perception, The Netherlands.

Sumner, C., Lopez-Poveda, E., O'Mard, L., and Meddis, R. (2002). A revised model of the inner cell and auditory-nerve complex. *Journal of Acoustical Society of America*, 111(5):2178–2188.

Sumner, C., Lopez-Poveda, E., O'Mard, L., and Meddis, R. (2003a). Adaptation in a revised inner-hair cell model. *Journal of Acoustical Society of America*, 113:893–901.

Sumner, C., O'Mard, L., Lopez-Poveda, E., and Meddis, R. (2003b). A nonlinear filter-bank model of the guinea pig cochlear nerve: Rate responses. *Journal of Acoustical Society of America*, 113:3264–3274.

Tan, Q. and Carney, L. (2003). A phenomenological model for the responses of auditory-nerve fibers. ii. nonlinear tuning with a frequency glide. *Journal of Acoustical Society of America*, 114(4):2007–2020.

Tan, Q. and Carney, L. (2005). Encoding of vowel/like sounds in the auditory nerve: Model predictions of discrimination performance. *Journal of Acoustical Society of America*, 117(3):1210–1222.

Tan, Q. and Carney, L. (2006). Predictions of formant-frequency discrimination in noise based on model auditory-nerve responses. *Journal of Acoustical Society of America*, 120(3):1435–1445.

Wang, H., Gelbart, D., Hirsch, H., and Hemmert, W. (2008). The value of auditory offset adaptation and appropriate acoustic modeling.

Westerman, L. and Smith, R. (1984). Adaptation and recovery of auditory nerve response. *Hearing Research*, 15:249–260.

Wiegrebe, L. and Meddis, R. (2004). The representation of periodic sounds in simulated sustained chopper units of the ventral cochlear nucleus. *Journal of Acoustical Society of America*, 115(3):1207–1208.

Winter, I. and Palmer, A. (1991). Intensity coding in low-frequency auditory nerve fibers of the guinea-pig. *Journal of Acoustical Society of America*, 90:1958–1967.

Zhang, X. and Carney, L. (2005). Analysis of models for the synapse between the inner hair cell and the auditory nerve. *Journal of Acoustical Society of America*, 118(3):1540–1553.

Zhang, X., Heinz, M., Bruce, I., and Carney, L. (2001). A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression. *Journal of Acoustical Society of America*, 190(2):648–670.

Zilany, M. and Bruce, I. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *Journal of Acoustical Society of America*, 120(3):1446–1466.

Zilany, M. and Bruce, I. (2007). Representation of the vowel /ε/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats. *Journal of Acoustical Society of America*, 122(1):402–417.

Zilany, M., Bruce, I., Nelson, P., and Carney, L. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *Journal of Acoustical Society of America*, 126(5):2390–2412.

Erklärung der Selbstständigkeit

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die Zitate deutlich kenntlich gemacht zu haben.

.....

Ort, Datum Nuria Vendrell Llopis