SOUND SOURCE CODING IN THE AZIMUTHAL PLANE: SEPARATING SOUNDS VIA SHORT-TERM INTERAURAL TIME DIFFERENCE ESTIMATIONS

Sebastian Groß

Dissertation der Graduate School of Systemic Neurosciences
der Ludwig-Maximilians-Universität München

June 15, 2021

Sebastian Groß : *Sound Source Coding in the Azimuthal Plane: Separating Sounds via Short-Term Interaural Time Difference Estimations*

Supervisor: Prof. Dr. Christian Leibold
Second reviewer: PD Dr. Michael Pecka
External reviewer: Prof. Dr. Mathias Dietz

Defense date: October 7, 2021

To my family and life-long friends.

CONTENTS

LIST OF ABBREVIATIONS

AIC	**A**kaike **I**nformation **C**riterion
AN	**A**uditory **N**erve
ANF	**A**uditory **N**erve **F**iber
AP	**A**ction **P**otential
AVCN	**A**nterio-**V**entral **C**ochlear **N**ucleus
BIC	**B**ayesian **I**nformation **C**riterion
BM	**B**asilar **M**embrane
BMLD	**B**inaural **M**asking **L**evel **D**ifference
BF	**B**est **F**requency
BP	**B**est **P**hase
CD	**C**haracteristic **D**elay
CF	**C**haracteristic **F**requency
CNS	**C**entral **N**ervous **S**ystem
CP	**C**haracteristic **P**hase
$\mathbb{C}$	**C**omplex numbers
DNLL	**D**orsal **N**ucleus of the **L**ateral **L**emniscus
EM	**E**xpectation-**M**aximization Algorithm
EPSP	**E**xcitatory **P**ost**s**ynaptic **P**otential
ERB	**E**quivalent **R**ectangular **B**andwidth
FCT	**F**ourier **C**onvolution **T**heorem
GBC	**G**lobular **B**ushy **C**ell
GMM	**G**aussian **M**ixture **M**odel
LSO	**L**ateral **S**uperior **O**live
lLSO	**l**ow-frequency part of the **L**ateral **S**uperior **O**live
LNTB	**L**ateral **N**ucleus of the **T**rapezoid **B**ody
MGN	**M**edial **G**eniculate **N**ucleus
MNTB	**M**edial **N**ucleus of the **T**rapezoid **B**ody
MPA	**M**ean **P**opulation **A**ngle
MSO	**M**edial **S**uperior **O**live
$\mathbb{N}$	**N**atural numbers
NL	**N**ucleus **L**aminaris
HCN	**H**yperpolarization-activated **C**yclic **N**ucleotide-gated cation channels

IC	**I**nferior **C**olliculus
IHC	**I**nner **H**air **C**ell
ILD	**I**nteraural **L**evel **D**ifference
IPSP	**I**nhibitory **P**ost**s**ynaptic **P**otential
IPD	**I**nteraural **P**hase **D**ifference
ITD	**I**nteraural **T**ime **D**ifference
JND	**J**ust **N**oticeable **D**ifference
Kv1	Low-threshold voltage-gated potassium channels
$\mathbb{R}$	**R**eal numbers
RMSE	**R**oot-**M**ean-**S**quare **E**rror
SBC	**S**pherical **B**ushy **C**ell
SOC	**S**uperior **O**livary **C**omplex
SPL	**S**ound **P**ressure **L**evel
SPN	**S**uperior **P**araolivary **N**ucleus

ABSTRACT

The interaural time difference (ITD) is the main cue to perform sound localization for low-frequency sounds (below $\sim 2\text{kHz}$) in the azimuthal plane. The extractors for this cue are neurons of two nuclei of the mammalian auditory brainstem, the medial superior olive (MSO) and the low-frequency limb of the lateral superior olive (lLSO). The read-out mechanism on a population level is unknown as single neurons show different responses for frequency-varying stimuli. This poses a challenge especially for natural sound stimuli and complex auditory scenes which cover a wide range of frequencies, i.e., they have a very broad spectrum.

To find an encoder of ITDs, we have developed so-called effective population models of the human MSO and lLSO. They are effective in the sense that the individual neurons are each identified by their three defining properties which determine their frequency-dependent ITD tuning: the best frequency (BF), the characteristic delay (CD) and the characteristic phase (CP).

We have formulated an ITD decoding strategy in the 2d-space spanned by the membrane potentials of lLSO vs. MSO. From each hemisphere, a separate ITD can be decoded. These two estimations can be weighted and balanced accordingly to retrieve the location of sound sources in the horizontal plane. To this end, we make use of so-called short-term ITDs which are successive estimates in small time windows. Our results indicate that sound localization can be performed correctly in time windows as short as up to 1ms. To perform sound separation of stimuli within complex auditory scenes, we fit Gaussian Mixture Models to the short-term ITD estimate distributions. The results show that sound separation can be performed reliably when the long-term ITD estimation (which is a distribution of short-term ITDs) is made up of a time interval that is larger than 1s. Furthermore, we conclude that sounds can be separated and

reconstructed from complex auditory scenes solely based on one auditory cue, the ITD. [0]

[0] Disclaimer. Some of the material and results in this work have been presented in the form of an abstract and/or as part of a scientific poster; cf. References and Publication List.

1 INTRODUCTION

Where more is meant than meets the ear.

— John Milton

1.1 FUNDAMENTALS OF SOUND PROCESSING IN THE AUDITORY SYSTEM

1.1.1 *Making Sense of Sound*

Sound is produced by variations in air pressure. These variations are the result of objects moving towards or from patches of air. Objects moving towards a patch of air compress the air molecules herein making the air more dense (high air pressure). Objects moving away, in turn, rarefy the air molecules making the air less dense (low air pressure). These audible variations in air pressure over distance constitute a pressure wave (Bear et al., 2016). And thus a classic definition of *sound* is *a pressure wave that propagates through air* (Schnupp et al., 2011). This definition is not complete as it does not take into account that sound waves can also propagate through other mediums such as liquids, other gases and also solids. A more broad definition of sound is provided by the Merriam-Webster dictionary as *mechanical radiant energy that is transmitted by longitudinal pressure waves in a material medium (such as air) and is the objective cause of hearing*. For clear terminology, we will use the word sound and all its compounds (e.g., sound source, sound wave, etc.) when referring to the physical phenomena of hearing and the adjective *auditory* to all events concerning the perception of sound (Blauert, 1997).
Now consider that there are multiple sound sources of sound in a room. Then each sound source will produce individual pressure waves that propagate at the speed of sound (which is

approximately $343 m/sec$ or $767 mph$ at normal room temperature) in all directions (if the sound source is not directional). The ripples in the air will bounce off the floor, the ceiling, the walls and soon the air of the room will be filled with an extremely complex pattern of air pressure ripples. Nevertheless, our brain can successfully make sense of these sounds and extract various information (Schnupp et al., 2011). It can determine if the sound source is animate (e.g., a person talking, singing, humming, whispering or an animal hissing, purring, barking) or an inanimate object (e.g., a bell ringing, a bottle that is shattered on the floor, a book that falls out of the shelf, a broom scraping across the floor). It can also conclude where the different sounds are coming from in the room, i.e., localize the sound sources in the azimuthal plane.

In order to understand how the human auditory system can respond in such a remarkable way to simple variations in air pressure, we briefly overview the fundamentals of sound processing. To this end we first describe how the mechanical power of a sound wave entering the ear can elicit an electrical neural response in the auditory nerve (AN). Then we discuss the representation of frequency in the nervous system (tonotopy and phase locking). Finally, we look at how the neural responses of the AN travel further through the ascending auditory pathway up to inferior colliculus (IC) with an emphasis on binaural processing stages. Since the main focus of this thesis deals with the (binaural processing) nuclei of the medial superior olive (MSO) and the lateral superior olive (LSO), we accentuate those stages of sound processing where the sound information from both ears is combined.

1.1.2 *From Sounds to Neural Signals*

When a sound is emitted, the first component of the outer ear the sound waves reach is the pinna (see Figure 1.1). Together with cartilage it is the only visible part of the outer ear and acts as a funnel with which humans can gather sounds. The shape of the pinna with its specific folds also helps to distinguish if a sound is coming from the front or back (azimuth) or from up or down (elevation). When the sound waves reach the outer ear, they further travel through the auditory canal and cause vibrations in the membrane of the tympanum. This vibration is passed on to the three ossicles of the middle ear: the malleus, incus and stapes. The footplate of the stapes pushes against the membrane of the oval window, a small opening in the skull where the inner ear begins. Whereas outer and middle ear are air-filled, the cochlea of the inner ear is fluid-filled. To achieve sufficient movement of the cochlear fluid, the auditory system makes use of two sound amplifying principles. Firstly, the three middle ear components play a crucial role in amplification, because the energy of a sound wave would be almost fully attenuated by the pressure of the cochlear fluid at the oval window. The ossicles function as levers and transform the large movements at the tympanic membrane into smaller albeit stronger movements at the oval window. A second source of sound amplification is that the surface area of the oval window is much smaller than of the tympanic membrane, i.e., the force arriving at the oval window is much higher than at the tympanic membrane. Thus, the tympanum and the ossicles act as an impedance-matching device for the difference of sound impedance at the outer and the inner ear. The movement at the oval window then sets the fluid of the cochlea into motion. Running throughout the length of the cochlea is the basilar membrane (BM) which divides the cochlea into two upper (scala media, scala vestibuli) and one lower compartment (scala tympani). The Organ of Corti – the site of signal transduction – is an epithelium within the scala media and is directly located on the BM. For signal transduction, the most important cells are the inner hair cells (IHC) of the Organ of Corti. When the BM is deflected by the movement of the fluid in the cochlea, the sterocilia located on top of the IHCs are displaced which opens mechano-gated channels within the cilia. An influx of potassium (K^+) ions leads to a depolarization of the IHCs which opens voltage-gated calcium (Ca^{2+}) channels. The subsequent influx of calcium leads to the fusion of vesicles filled with excitatory neurotransmitters (glutamate) with the presynaptic membrane of the IHC. After release into the synaptic cleft, the neurotransmitters diffuse to the postsynaptic spiral ganglion neurons. The AN is made up of all axons of the neurons whose cell bodies are within in the spiral ganglion. Action potentials (APs) are then generated at the soma of each spiral ganglion cell. This is the first stage of the auditory pathway where APs are fired. Thus the mechanical energy of sound through movement of air molecules has been transduced by the IHCs into an electrical neural signal which is transported via these APs to subsequent nuclei of the ascending auditory

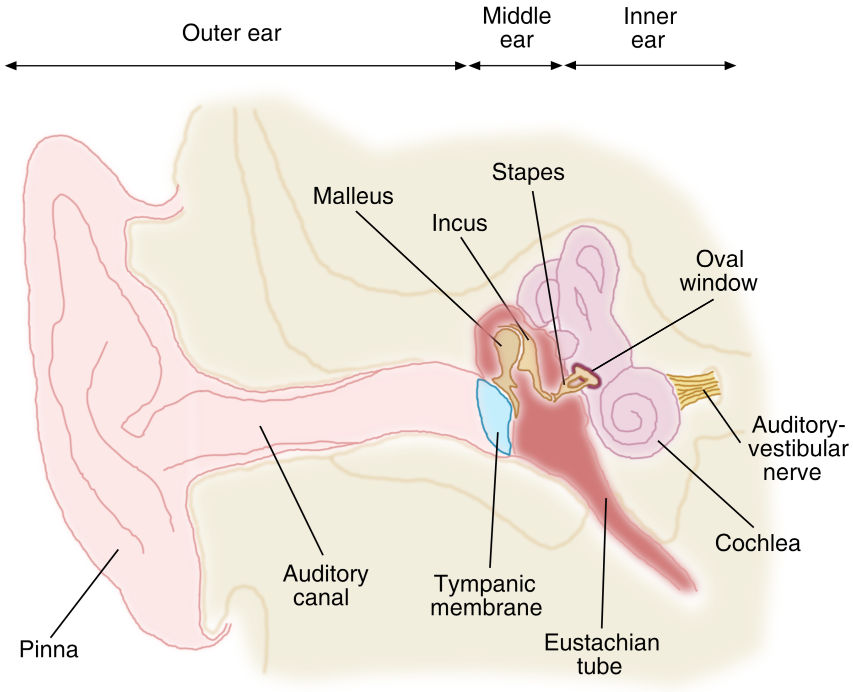

Figure 1.1: Cross section of human ear. See main text for further explanation. Figure inspired by Bear et al. (2016).

pathway (see Section 1.1.4) where the incoming information is then further processed (Zenner, 1994; Schnupp et al., 2011; Brown and Santos-Sacchi, 2012; Kandel, et al., 2013; Luo, 2015; Bear et al., 2016).

1.1.3 *Frequency Representation in the Auditory System*

Since neurons of the auditory system show frequency-dependent varying responses (see Section 1.2.2), one important question is how frequency is represented in the auditory system. Two mechanisms are tonotopy and phase locking.

Tonotopy. Tonotopy is the organization of frequency within an auditory structure (von Békésy and Wever, 1960). The BM features tonotopy as a direct consequence of its anatomical structure (see Figure 1.2A). The base of the BM is narrow and stiff, the apex is wide and flexible. The movement at the membrane of the oval window in response to sounds sets the cochlear fluid (perilymph) into motion bending the BM at the base. This sets a traveling wave along the BM into motion. The frequency of a sound determines how far this wave travels along the BM. High-frequency sounds cause strong vibrations at the base. These vibrations cause much of the energy of the wave to be dissipated at this point and therefore the wave will not travel far from the base. In turn, low-frequency sounds do not cause very strong vibrations at the base and therefore much of the energy is not dissipated. This can generate waves that can travel to the end of the BM, the apex. In sum, sounds with differing frequencies will maximally displace the BM at different positions and thus frequency is represented on the BM via a place code. This tonotopy is preserved throughout the ascending auditory pathway up to auditory cortex (Schnupp et al., 2011; Brown and Santos-Sacchi, 2012; Bear et al., 2016).

Phase Locking. AN fibers can also represent the frequency of pure tones by locking their action potential response to a specific sound phase (Galambos and Davis, 1943; also see Figure 1.2B). A pressure wave generated via a pure tone will repeat itself after one whole cycle (or period). A neuron which is locked to a certain phase will always fire an action potential whenever a specific phase of the sound is present and only when this phase is present. Thus the frequency of the neuron's firing directly reflects the frequency of the sound (Joris et al., 1994). Phase locking breaks down for sounds with high frequencies above 4kHz because the cycles become too short to be reliably represented by AP firing (an AP takes approximately $1ms$). For sounds above 4kHz, frequency is then solely represented by tonotopy (Zenner, 1994; Luo, 2015; Bear et al., 2016). It is important to note at this point, that phase locking is not only a means of frequency representation in the auditory system, but furthermore a prerequisite for the functioning of fast temporal processing in the downstream MSO and lLSO neurons (see Discussion, 3.3.1).

1.1.4 *Superior Olivary Complex (SOC) Circuits in the Ascending Auditory Pathway*

Until sound reaches the final stage of sound processing – the auditory cortex – the APs generated in the AN are relayed via several auditory structures which, as a whole, are referred to as the ascending auditory pathway. For reasons of conciseness we

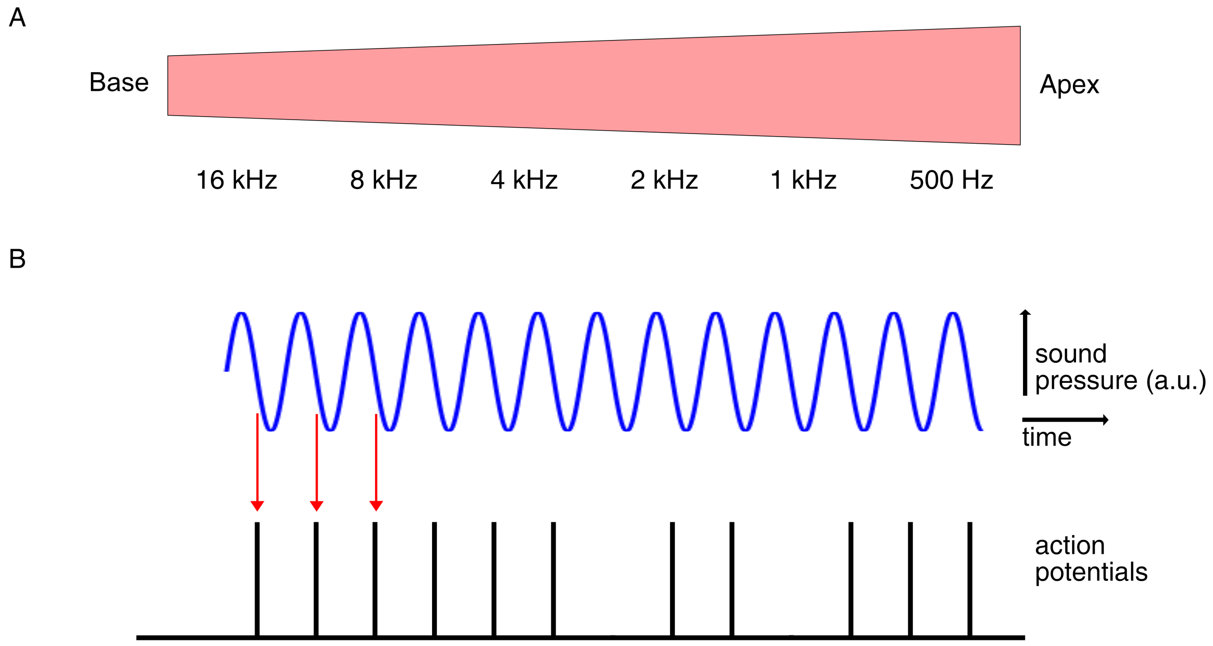

Figure 1.2: Frequency representation. A: Structure of the BM within an uncoiled cochlea. Sounds of different frequencies maximally displace the BM at different positions. Frequency is thus represented via a place code along the entire BM (tonotopy). B: Schematic of phase-locking. A low-frequency pure tone (top graph) evokes a response in form of an action potential (black vertical lines, bottom graph) in an AN neuron whenever the sound is at a specific phase (red arrows). The neuron only shows response to exactly this phase on (almost) every cycle of the sound, i.e., it is locked to this phase. Note that it need not be the case that the neuron fires on every cycle. Figure inspired by Bear et al. (2016) and Lehnert (2015).

mainly focus on the inputs and outputs to the MSO and LSO as these are the first two structures in auditory brainstem which integrate and evaluate information from both ears, i.e., they are the first two nuclei where binaural processing takes place.

MSO Circuit. The MSO is a nucleus within the SOC which is a collection of nuclei located in the auditory brainstem (overview in: Grothe et al., 2010 and Yin et al., 2019; Figure 1.3). The SOC consists of three further nuclei, namely the LSO as well as the

[1] Physiological Reviews, *Mechanisms of sound localization in mammals*, Benedikt Grothe, Michael Pecka, David McAlpine, VOL 90, July 2010, p. 995.

lateral and medial nucleus of the trapezoid body (LNTB and MNTB, respectively). MSO principal neurons exhibit a typical bipolar shape that receive binaural glutamatergic excitatory and binaural glycingergic inhibitory inputs. Since MSO cells are predominantly binaurally excited, the MSO cells are described as excitatory-excitatory (EE) neurons (Goldberg and Brown, 1969). Sources of excitation are the ipsilateral and contralateral spherical bushy cells (SBCs). The SBCs together with another group of cells, the globular bushy cells (GBCs), are subgroups of the cells of the anterio-ventral cochlear nucleus (AVCN). The SBCs and GBSs themselves directly receive input from the AN fibers, i.e., they are the first relay stage after signal transduction in the central nervous system (CNS). Sources of inhibition to the MSO are cells of the ipsilateral LNTB and MNTB. These two nuclei are innervated from two different sides: The LNTB itself

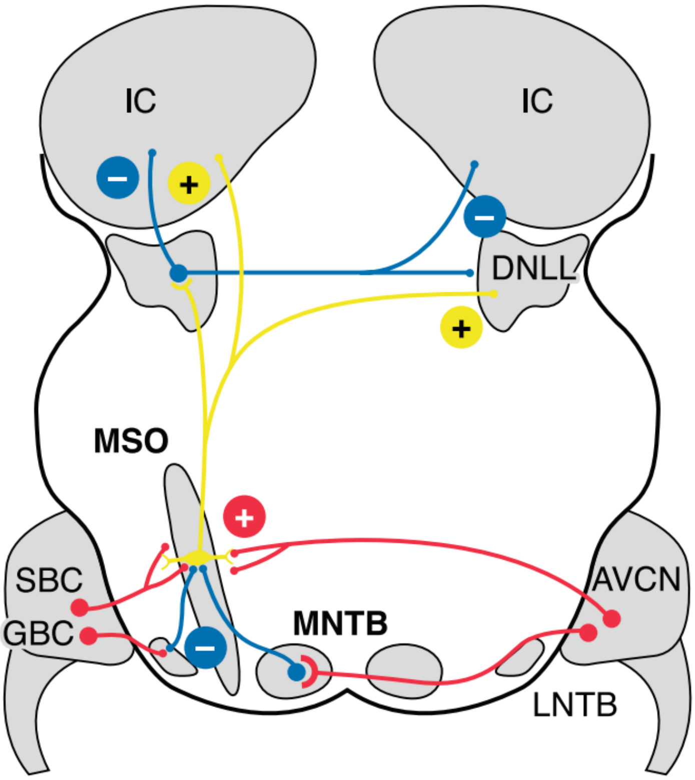

Figure 1.3: Upstream and downstream circuitry of the medial superior olive. Color Code: Blue = Glycinergic inhibition. Red = Glutamatergic excitation. Yellow = Typical MSO neuron with its glutamatergic excitatory projections. See main text for full description. Figure reprinted with permission from Grothe et al. (2010)[1].

receives excitation from the ipsilateral GBCs of the AVCN and the MNTB itself receives excitation from the GBCs of the contralateral AVCN. At this point it may be already mentioned, that the cells of the MNTB receive their input through the largest synapse of the brain, the Calyx of Held. Excitation of the MNTB therefore occurs in such a fast manner that the phase-locked response of the AN is preserved, guaranteeing temporally precise phase-locked inhibition to the MSO. After integration of excitatory and inhibitory inputs, the MSO itself sends excitatory projections to the ipsi- and contralateral dorsal nucleus of the lateral lemniscus (DNLL) and ipsilateral inferior colliculus (IC). The lateral lemniscus is an axon bundle that leads to the IC of the midbrain. Neurons of the IC then project to auditory thalamus, more precise the medial geniculate nucleus (MGN), which finally projects to the last stage of the ascending auditory pathway, the auditory cortex.

LSO Circuit. The LSO is a nucleus in the SOC which is located laterally with respect to the MSO (overview in: Grothe et al., 2010 and and Yin et al., 2019; Figure 1.4). In comparison to the MSO, the LSO only receives one ipsilateral excitatory and one contralateral inhibitory input. The cells of the LSO are therefore described as excitatory-inhibitory (EI) neurons (Goldberg and Brown, 1969). The SBCs of the ipsilateral AVCN provide glutamatergic excitation, the ipsilateral MNTB provides glycinergic inhibition. The MNTB is innervated contralaterally and is excited (also through the Calyx of Held synapse) by the GBCs of the contralateral AVCN. The LSO sends excitatory projections to contralateral DNLL and IC and inhibitory projections to ipsilateral DNLL. Analogously to the MSO circuit, the information is then passed on to higher auditory regions via MGN and finally auditory cortex. Unique to the LSO circuit, is a de novo interaural level difference (ILD, also see Section 1.2) sensitivity which is generated due to converging inputs, a monaural input from the contralateral AVCN (excitatory) and a binaural input from the DNLL (inhibitory). It is noteworthy at this point, that the MSO and LSO circuit not only have similar design principles, but that they also share circuit components. A key notion that hints at the fact that the information of MSO and LSO may in fact not be processed completely independent of each other, but

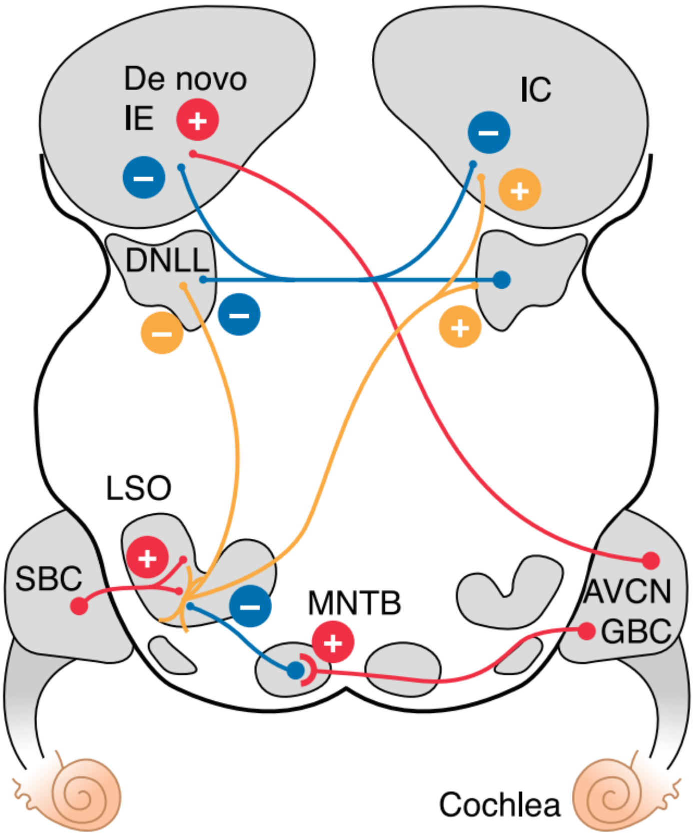

Figure 1.4: Upstream and downstream circuitry of the lateral superior olive. Color Code: Blue = Glycinergic inhibition. Red = Glutamatergic excitation. Orange = LSO neuron with its projections. See main text for full description. Figure reprinted with permission from Grothe et al. (2010)[2].

that information from both nuclei may be combined at higher auditory stages (see Section 1.4.1).

1.2 SOUND LOCALIZATION IN THE AZIMUTHAL PLANE

1.2.1 *Cues for Azimuthal Sound Localization: ITDs and ILDs*

There are two dominant cues for sound localization in the azimuthal (or horizontal) plane, the interaural time difference (ITD) and the interaural level difference (ILD).

[2] Physiological Reviews, *Mechanisms of sound localization in mammals*, Benedikt Grothe, Michael Pecka, David McAlpine, VOL 90, July 2010, p. 991.

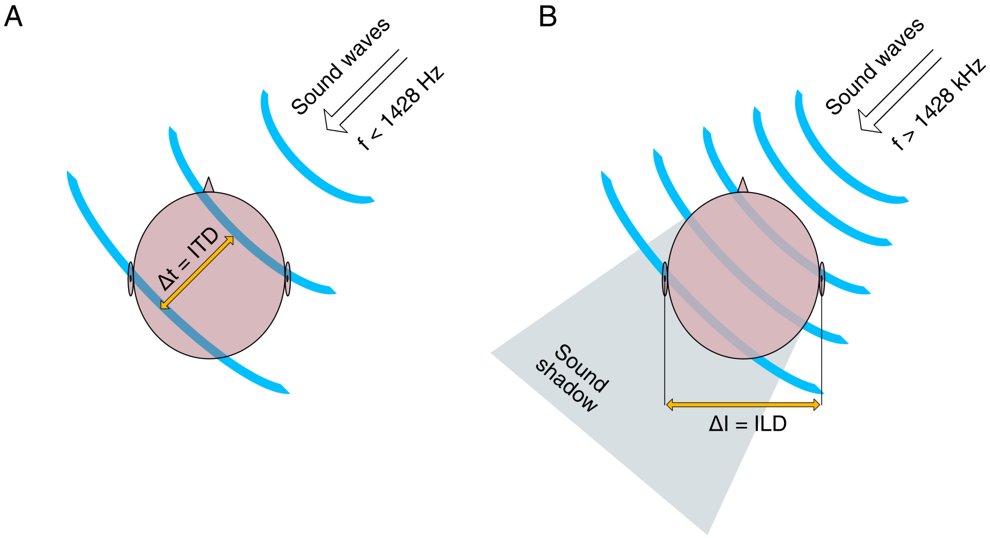

Figure 1.5: Cues for azimuthal sound localization. A: Low-frequency sounds are located via ITDs, the interaural time difference (denoted by Δt). B: High-frequency sounds are located via ILDs, the interaural level or intensity difference (denoted by Δl). Making use of a specific cues depending on the frequency of the sound is referred to as the *duplex theory of sound localization*.

ITDs. When a sound is coming from the right, it will arrive earlier at the right ear than at the left ear (Figure 1.5A). This difference in arrival time is referred to as the ITD (review: Grothe et al., 2010). Conversely, if a sound is coming from the left, it will arrive at the left ear earlier than at the right. A sound coming from straight in front (or behind) the head will arrive at both ears simultaneously. For convention, we denote sounds coming from the right with positive ITDs and sounds coming from the left with negative ITDs. An ITD of zero represents a sound source position at midline. The maximal possible difference in arrival time is bounded by the head size (more precisely the ear-to-ear distance) of the species in question and is referred to as the *physiological range*. For example, for gerbils (*Meriones unguiculatus*) this range is approximately $\text{ITD} = \pm 135\mu\text{s}$ (Maki and Furukawa, 2005) and for humans it is approximately $\text{ITD} = \pm 700\mu\text{s}$ (Kuhn, 1977), where the negative sign corresponds to ipsileading and the positive sign refers to contraleading sounds. For humans, ITDs can only be used as a localization cue if the sound contains frequencies below the inverse of $2 \cdot 700\mu s$, namely $\sim 1428\text{Hz}$ as the wavelength of the incoming sound must be larger than the distance between the ears. If the wavelength is shorter than this distance, then ITDs cannot be revolved unambiguously as it becomes impossible to distinguish individual cycles of the incoming sound. The main extractor for ITDs is the MSO, but the LSO also shows sensitivity to ITDs. The neurons of the LSO most responsive to low frequencies are located in the lateral limb (Tollin and Yin, 2005) and we denote this by lLSO (low-frequency limb of the LSO).

ILDs. For high-frequency sounds above $\sim 1428\text{Hz}$, another cue is utilized to locate sounds in the horizontal plane. This cue makes use of the level or intensity of a sound. When a sound is coming from the right, the sound will have a higher level at the right ear than at the left ear. This level difference is referred to as the ILD (review: Grothe et al., 2010). Conversely, if a sound is coming from the left, it will have a higher level at the left than at the right ear. A sound coming from straight in front (or behind) the head will result in identical intensity at both ears. The difference in level is a direct consequence of the sound shadow that high-frequency sounds cast around the head. Sound shadows occur because the incoming sounds are directly reflected by the head resulting in a lower intensity at that ear that is within the sound shadow. Note that for low-frequency sounds ($\lesssim 2\text{kHz}$) there is almost no sound shadow because the sounds are not reflected by the head, they are rather diffracted around the head resulting in identical intensities arriving at both ears. The main extractor for ILDs is the LSO, but it is also known that the MSO has neurons that are sensitive to ILDs.

Summing up the two processes, we have two different cues for localizing sounds in the horizontal plane. For low-frequency sounds ($\lesssim 2\text{kHz}$) we make use of ITDs and for high-frequency sounds ($\gtrsim 2\text{kHz}$) we make use of ILDs. The separation of sound localization cues into two frequency bands is called the *duplex theory of sound localization* (Rayleigh, 1907).

1.2.2 *Neural Representation of Acoustic Space: The Jeffress Model*

The question we now focus on is how can acoustic space be neurally represented, i.e., how is sound source coded into the brain. One famous theoretical model was presented by Jeffress in 1948 where he proposed that there are binaural coincidence detector neurons located in the MSO which are sensitive to different ITDs due to so-called delay lines (Jeffress, 1948; review: Leibold and Grothe, 2015). The idea is that neurons in the MSO are lined up next to each other in a row (see Figure A, right). Every neuron receives purely excitatory inputs from both AVCNs (binaural). When a sound, e.g., is coming from the right, it will trigger an earlier response in the right AVCN than in the left AVCN. There now exists one neuron in the MSO which responds maximally to this difference in timing because there are delay lines (afferents) leading to it and their lengths are varied in such a way that the responses from both sides will arrive at the same time at this neuron, hence the name coincidence detector neurons. The afferents to the neuron in this example will therefore be longer from the right AVCN than from left AVCN to cause this compensation of delays.[3] The Jeffress model thus states that for each ITD there exists a different neuron that will respond maximally to this ITD by delay line compensation from both AVCNs as compared to neurons surrounding it. This results in the concept that different neurons are tuned to a specific ITD, their best ITD. This type of neural representation suggests a peak code where the ITD is encoded by the maximum firing rate of neurons, i.e., the firing rates of all neurons in one MSO are compared with each other and the highest firing rate yields the winner revealing the sound source position (see Figure B, left). The Jeffress model is solely based on conduction times and therefore it does not incorporate varying frequencies of sounds. Therefore a consequence is to expect a neuron to have the same maximum response to its tuned best ITD independent of frequency (see Figure B, middle and right).

We now focus on evidence which speaks against the implementation of the Jeffress model as the mechanism of neural

[3] It is important to note, that Jeffress did not write about the MSO or AVCN as he considered the delay lines to be located in IC. The model's general mechanism has nowadays been applied to the nuclei that fit today's understanding of the anatomy of the mammalian auditory brainstem.

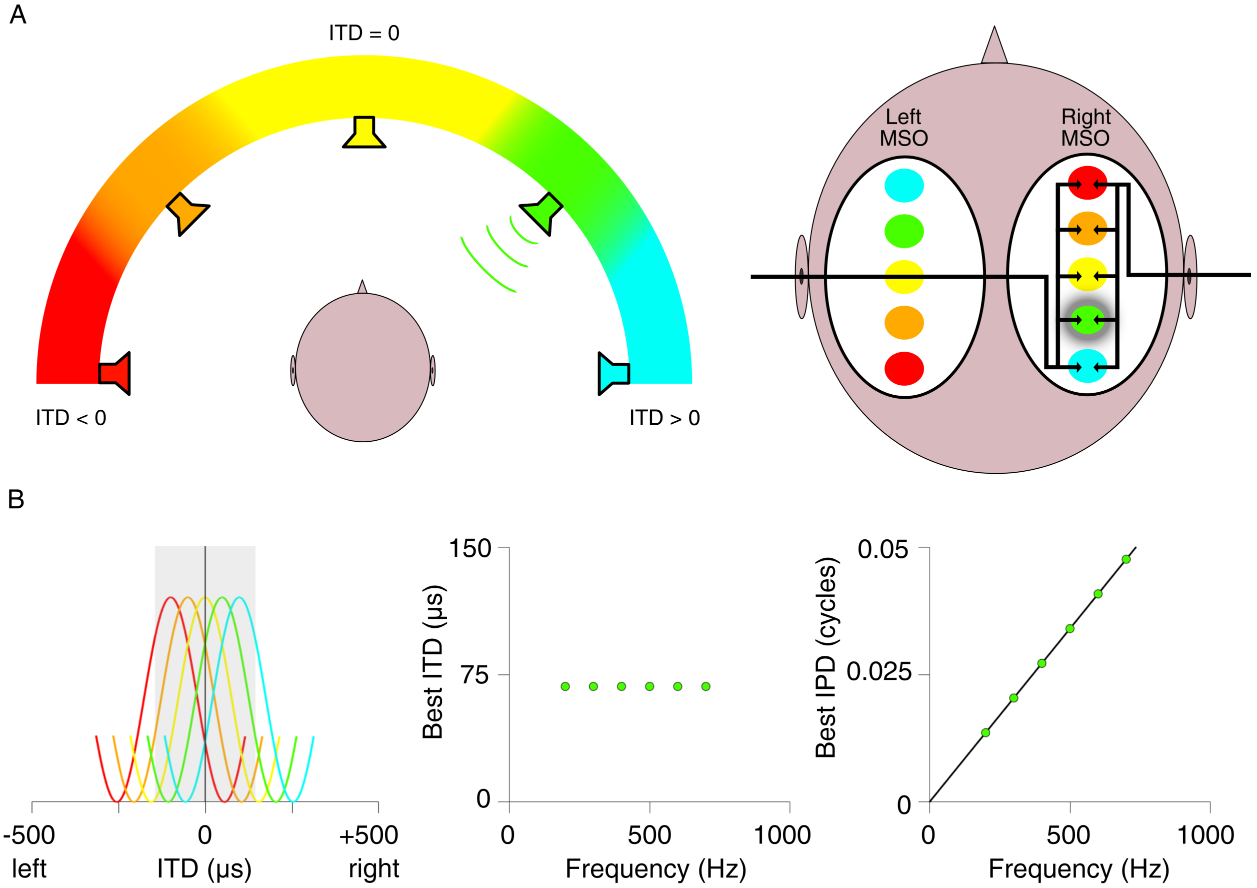

Figure 1.6: The Jeffress model. A: If the green speaker emits an acoustic signal, the sound waves reach the right ear earlier than the left ear, leading to an earlier response in the right AVCN than the left. Action potentials travel down afferents, called delay lines. The length of these delay lines are varied in such a way, that the responses from both ears maximally coincide at the green neuron (colored disk). The delay line from the right is thus longer than from the left AVCN. This results in a maximal firing rate for each neuron to their specific tuned ITD, their best ITD. B, left: Schema of tuning curves (firing rate vs. ITD). Every neuron (different colors) has its maximum firing rate to one fixed ITD establishing a peak code. Note that all peaks are within the physiological range (gray shaded area). B, middle: Best ITD vs. sound frequency. The maximal response from one neuron to its best ITD is independent of frequency. B, right: Multiplying the Best ITD with sound frequency results in the interaural phase difference (IPD). Best IPD vs. sound frequency is a linear function (straight line) with slope corresponding to the best ITD. Frequency-invariance is geometrically represented by the intercept of the line at best IPD $= 0$ cycles. Figure inspired by Leibold and Grothe (2015).

representation of auditory space in mammals. One reason that speaks against it, is the non-existence of systematic delay lines. Even though in the nucleus laminaris (NL), the avian analogue of the MSO, such delay lines have been reported (Carr and Konishi, 1990; Seidl et al., 2010), for mammals they are absent as 3D reconstructions (Karino et al., 2011) and *in vivo* recordings have shown (Franken et al., 2015). Also the purely excitatory neural circuit proposed by Jeffress does not fit the anatomy as described in 1.1.4. The MSO not only receives two glutamatergic excitatory inputs (from both AVCNs) but also two glycinergic inhibitory inputs from (LNTB and MNTB). Furthermore, for the Jeffress model to hold true, the firing rate peaks describing the best ITD of each neuron would have to be spread (homogeneously) within the physiological range of the animal. But studies have shown (e.g. McAlpine et al., 2001) that neurons rather have their maximum firing rate outside the physiological range (see Figure 1.7, left), creating a strictly monotonous part with a steep slope inside the physiological range from which ITDs could be inferred, hinting at a rate code rather than a peak code. And further, if glycinergic inhibition was blocked, then the peaks of several neurons shifted to zero making an ITD readout impossible and thus negating the purely excitatory Jeffress mechanism.

Physiologically, the most important quantity which does not fit the structure of the Jeffress model is the characteristic phase (CP), the frequency-dependent delay difference from both ears. As mentioned earlier, the peak firing rate of a neuron in Jeffress' model is determined by a conduction delay difference from the incoming axons from both ears and such a delay difference would be independent of sound frequency (see Figure B, middle). Such a frequency-independent delay difference is called characteristic delay (CD). However, there is experimental data (Day and Semple, 2011; Franken, 2015) which suggests that the Best ITD changes for a single neuron when the frequency of the stimulus is varied (see Figure 1.7, middle and right), showing that there must be a frequency-dependent delay difference from both ears, coined the CP. Nevertheless, the existence of CPs and their broad distributions across animal species make it a fundamental quantity which must be incorporated into a model of neural representation of auditory space to make it a biologically feasible one.

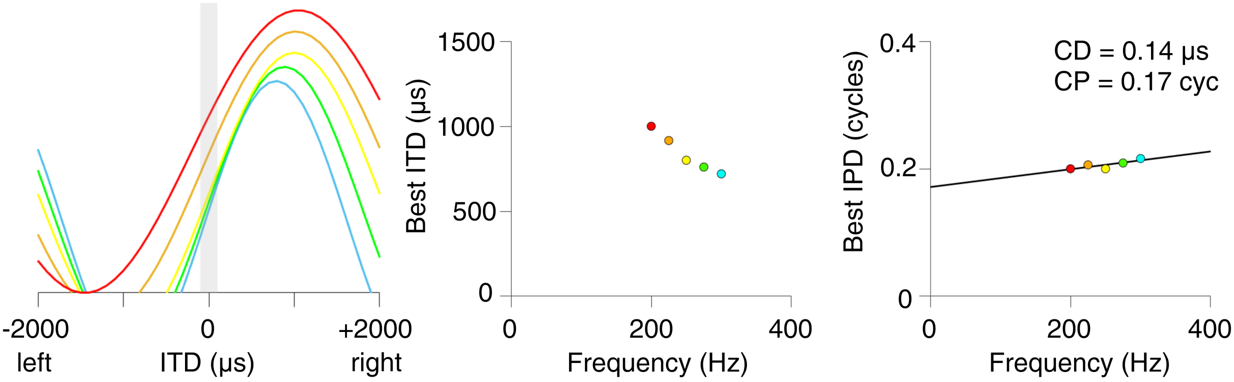

Figure 1.7: Characteristic phases. Left: Firing rate vs. ITD. The tuning curves show a schematic response of a single gerbil MSO neuron to sinusoidal sounds. Varying colors represent different frequencies. Peaks indicate the Best ITD. Note that the peaks are all located outside the physiological range (gray shaded area) with a contralateral preference. This establishes strictly monotonous parts with steep slopes within the physiological range which establishes a rate code rather than a peak code. ITDs are encoded best by steep slopes since for a fixed ITD the firing rates of neighbouring ITDs will show a large difference. Middle: Best ITD vs. sound frequency. The best ITD varies with stimulus frequency which is in opposition to the Jeffress model. Right: Best IPD vs. sound frequency and linear function fit (straight line obtained via linear regression) with slope corresponding to the best ITD. Frequency-variance is geometrically represented by the non-vanishing intercept of the line at Best IPD $= 0.17$ cycles. Figure inspired by Leibold and Grothe (2015).

1.3 Models of Binaural Hearing: An Overview

We now turn to a brief overview of the history of a variety of models that can each be interpreted as models of binaural hearing, which can estimate the azimuthal position of a sound source. The models have been chosen to cover a wide variety of model types and therefore a special focus in this section will be on the motivation behind these models and if the models in question can be regarded as accurate representations of the underlying biology.

1.3.1 *Left-Right Comparison Models*

Left-right comparison models are lateralization models which compare counts or activity from hypothetical left and right neural populations. Each population makes up a so-called channel. The information from the channels' activity can then be used to determine where a sound is located in the azimuthal plane (Colburn and Durlach, 1978).

Tuning Models. The first type of this model dates back to 1930 and was suggested by von Békésy. In the model, von Békésy (1930) defines a group of cells that are centrally located. All cells receive inputs from the left and right ear. A neuron can enter one of two states, a left-tuned or a right-tuned state. If a stimulus is located on the right, it will reach the right ear before the left one. Whenever a stimulus reaches an ear, it sends a wave of excitation at a constant rate across the cells. Depending on which wave (from the left or right) reaches a neuron first, this neuron will then enter a left- or right-tuned state. When the waves from both ears meet, they extinguish each other and tuning is complete. The position of the stimulus source is then obtained by counting the relative number of left-tuned and right-tuned cells. This model was an explanation of the observation of the shift of virtual (acoustic) images in lateralization experiments. For small ITDs, this shift was very large as compared to higher ITDs. He explains this phenomenon by an increased density of cells in the middle as compared to the sides. One important conclusion from this is that around midline neurons should

have their highest sensitivity to varying ITDs and thus, when considering tuning curves, the maximum of these curves cannot lie at midline since steep slopes resemble high ITD-sensitivity (also see *Further Left-Right Comparison Models* below). The model from von Békésy is purely motivated by psychophysics and with our modern understanding of neurons, it has no biological foundation, since von Békésy's model implies that neurons at higher stages are able to distinguish between two types of firing: a response either received from a left or a right wave of excitation. But neurons cannot respond in more than two ways, since they respond in an all-or-nothing manner which means that they either give a response or they do not (Hall, 1964).
A modification of this model was provided by van Bergeijk (1962) in his *Variation on a Theme of Békésy*. He considers two different groups of neurons (whereas von Békésy only considered one group) and each group receives contralateral excitation as well as ipsilateral inhibition. Similar to von Békésy's excitation wave mechanism, this leads to areas of excitation and inhibition in both groups, i.e., the number of cells excited or inhibited. Higher auditory centers would then compare the areas of excitation from the left and the right group. This model, as an extension of the Békésy model, is also of course psychophysically motivated but it is also biologically motivated to the extent that it incorporates two "accessory nuclei of the superior olive" (van Bergeijk, 1962), which are the left and right MSO. (Hall, 1964; Colburn and Durlach, 1978; Stecker and Gallun, 2012)

Equalization and Cancellation Model. In 1963, Durlach used a signal processing approach to describe observations in binaural detection experiments. His idea comes from the fact that subjects can detect target signals in binaural-masking stimuli (which are comprised of a target and a masking noise signal) when the phase of target and masker are offset. This results in a reduction of detection threshold as compared to when phase of target and masker are equal, i.e., the target signal can more easily be unmasked. The difference between non-reduced and reduced threshold is referred to as the binaural masking level difference (BMLD). Durlach proposed that the auditory system performs an equalization (E) and a cancellation (C) process. In the E phase, the auditory system transforms the binaural-masking signal in one ear relative to the other ear until the masker is identical in both ears, i.e., it has been equalized. Then, in the C phase, the signal from one ear is subtracted from the other ear. If the auditory system would work in a completely precise manner, then now the masker would have been eliminated or cancelled (Durlach, 1963; Colburn and Kulkarni, 2005). Durlach states that the EC-theory can be used to explain how the auditory system can perform ITD discrimination. When turning off the target signal, then the listener could measure the ITD simply by recording the E-phase transformation and calculating its inverse which then would provide "a complete description of the interaural relations of the masking signal" (Durlach, 1963). Although the EC-theory is purely mathematical and based on signal processing, it still performs well in sound source separation algorithms (Mi et al., 2017).

Further Left-Right Comparison Models. McAlpine et al. (2001) took up the idea of left-right comparison models again. Recordings from IC of guinea pig revealed the peaks of ITD tuning curves to be located outside of the physiological range, rendering them unimportant for a sound localization (place) code. Furthermore, they showed that the steepest parts of the tuning curves were always located at midline and independent of BF. They therefore postulate a rate code for ITD estimation which is generated by "the activity in two broad, hemispheric spatial channels". (McAlpine et al., 2001). Motivationally speaking, the 2-channel model proposed is not based on biological or physiological properties since it is the result of recordings.
Another left-right comparison model is that of Dietz et al. (2008). It is a modified implementation of the excitatory-inhibitory (EI) coincidence detector model of Breebaart et al. (see below). In Dietz's model, the place code from Breebart is exchanged with an IPD rate code. The idea for this was initiated by the physiological findings in McAlpine et al. (2001) and Harper and McAlpine (2004). The model is primarily motivated by filter mechanisms and for the binaural processing to work, it relies on the precise extraction of the phase of an analytic signal *past* the stage of the hair cells. It is debatable and unknown if such a mechanism exists in the auditory system. Both of these reasons speak against a biologically motivated model albeit it is a phenomenological model of the auditory system. Another usage of left-right comparison models is Takanen et al. (2014) where they visualize the rate of output of the auditory system. Their idea is that the output of a comparison model does not directly result in a topographic map, so they present a method

of extracting binaural activity maps and further show, that the model performance agrees with human psychoacoustics.

1.3.2 *Coincidence Detection Models*

Models that are based on the underlying idea that an array of neurons register simultaneously arriving inputs from both ears are called coincidence detection models. Taking all neurons together, they display a wide range of internal delays and thus constitute a topographic map from which ITDs could be extracted. Such an encoding mechanism of ITDs is often referred to as a labeled line code, because each neuron along the array (line) is labeled for exactly one ITD. Groundwork for these types of models is the above discussed Jeffress model. As already mentioned, for mammals the Jeffress model seems unsuitable, since there is no anatomical evidence for the delay lines he proposed. Regardless, there exists a wealth of Jeffress-like model implementations, e.g., Colburn (1973, 1977), Lindemann (1986), Stern and Colburn (1978), Stern et al. (1988) and Breebart et al. (2001). They all have in common that initial stages are of phenomenological nature, e.g., the implementation of auditory periphery via gammatone filterbank and subsequent nonlinear reconstructions of the envelope, but characteristics of cells are not taken into account at all. Rather, *all possible* delays that could theoretically occur are calculated and implemented via binaural cross-correlation (hence why these types of models are also referred to as Jeffress correlation models (Colburn and Kulkarni, 2005). With time, the models have shifted away from the purely excitatory mechanism proposed by Jeffress and instead have incorporated aspects of inhibition, e.g., Lindemann (1986) and Breebart et al. (2001) both include contralateral inhibition. It is even more prominent in Breebart's model where the two binaural inputs are compared with each other via an excitatory-inhibitory (EI) coincidence detector mechanism based on the Equalization and Cancellation Model of Durlach. Even though all the models discussed in this section are successful at predicting the performance of human subjects in psychophysical tasks, we must regard these models as not biologically motivated for mammals even though the models do incorporate some important phenomenology of the mammalian auditory system (Stecker and Gallun, 2012). The most recent model using a coincidence detector is Klug et al. (2020), where they optimize the EI coincidence detector to account for lateralization in high-frequency stimuli. They show that the output of an EI coincidence detector in form of a hemispheric rate difference relates "linearly with the extent of laterality in human listeners" (Klug et al., 2020) and show this holds true for a thousand varying amplitude-modulated stimuli.

1.3.3 *Population Pattern Models*

Models based on machine learning algorithms that themselves find patterns in hemispheric activity are called population pattern models. The theory is, that a sound emitted at a certain location in azimuthal space will trigger a specific and unique pattern of activity over the whole population of neurons and if the position of the sound is shifted horizontally, then this will evoke a different unique pattern of activity across all neurons (Day and Delgutte, 2013). Population pattern models make use of a variety of different readout mechanisms, such as linear classifiers, rate difference decoders (Lüling et al., 2011), maximum-likelihood decoders (Day and Delgutte, 2013) or pattern match decoders (Goodman et al., 2013). Since these readout mechanisms themselves learn the patterns from the overall population activity, the way in which the activity from both hemispheres is used to extract relevant information (such as ITD information) is unknown *before* the decoder is applied. This results in the fact that population pattern codes can be very different from each other (depending on the readout mechanism applied) and also that these codes are very general, i.e., one code class can obviously contain other subsets of codes such as the previously mentioned left-right comparison code. The codes derived through population pattern models elude experimental falsification through their abstractness which, of course, make them seductively appealing in solving aspects which other models cannot, e.g., the non-uniqueness of ITD encoding for large heads due to the periodicity of ITD tuning curves (Leibold and Grothe, 2015; also cf. Discussion). Finally, these types of models do not resemble a 1:1 biological correspondence. It is one of the goals of our model (cf. next section) to aim at a fully biologically motivated representation of the MSO and LSO.

1.3.4 *Physiologically-Based Models*

Other models trying to explain the underlying ITD localization mechanisms and that are more physiologically grounded are, e.g., Encke and Hemmert (2018) where a spiking neuron network model of the MSO is used to extract ITDs in two different ways. In the first method, ITD changes are detected by a linear opponent channel decoder. In the second method, an artifical neuronal network is employed to predict ITDs based directly on the spiking output of their MSO and auditory nerve fiber (ANF) model. Their results show that ITD changes can be detected up to $10\mu s$ and that the MSO population can encode static and time dependent ITDs covering a wide range of frequencies, including complex sounds. Another model using the cell properties of MSO and LSO cells in a linear membrane voltage model is Remme et al. (2014). Based on recorded cells from guinea pig, they implement their cells to include a capacitive current, a leak current and a resonant and amplifying ion current. Using their dynamics, they show, that the subthreshold resonance properties of MSO and LSO cells can contribute to the efficient encoding of ITDs. The most noteworthy models with respect to this thesis are the physiologically-based models by Hancock and Delgutte (2004) and Hancock (2006) where a population rate code for ITDs is implemented via MSO and LSO by six parameters, the characteristic frequency (CF), a time constant τ, the CD, the CP and the coefficients A and B of a quadratic function which transforms the output of a cross-correlator into a firing rate. They retrieve the values by fitting the model to IC cat data (Hancock and Delgutte, 2004). They show for an $\text{ITD} = +1.5ms$ (not μs) that narrowband noise at 500Hz is perceived on the left and broadband noise $300 - 700\text{Hz}$ is perceived on the right. In general they show, that the broader the bandwidth of the input stimulus is, the closer the perception is to the ground-truth ITD position.

1.4 THE EFFECTIVE MODEL

The model presented in this thesis, which will in great detail be explained in Chapter 2, can be classified as belonging to the left-right comparison models, i.e., we use the activity from both hemispheres and combine the information in a meaningful way to derive a neuronal representation of azimuthal space. However, it is the first computational model which gives a full representation of the MSO and LSO nuclei according to their individual cell characteristics as have been reported *in vivo* (Tollin and Yin, 2005; Siveke et al., 2006; Siveke et al., 2007; Pecka et al., 2008; Lüling et al., 2011). Therefore, the model is novel and separates itself from the aforementioned ones, as it is a biologically and physiologically grounded model of sound localization. Due to its purely phenomenological description of the human binaural sound processing system, we thus term this model the *effective* model of binaural hearing.

Most similar to our effective model is the the model by Hancock (2006) as described in the previous section, but there are three important distinctions. Firstly, the CP and CD in their model is only applied to one side, the contralateral hemisphere. Secondly, the CP is set as a constant of $\mathrm{CP} = 0.5$ cycles for the LSO and lastly, the CP is set to a constant of $\mathrm{CP} = 0$ cycles for the MSO. In the effective model, we apply the CP and CD not only to both hemispheres, but we use a much higher variety for the values of the CP, making this model the first one to capture the biological reality of the characteristics of the mammalian MSO and LSO cells.

1.4.1 *Further Biological Motivation*

One of the key aspects of our model is that we not only use fully biologically based implementations of MSO and LSO neurons, but that we also combine the information from both nuclei. The motivation for this approach is the following. The descriptions of the MSO and LSO neuronal circuits in Section 1.1.4 are simultaneously a description of the mammalian ITD and ILD circuits, respectively. As previously mentioned, the two circuits not only share circuit components and have similar design principles, but both MSO and LSO act as detectors of binaural differences. The general idea is therefore that MSO and LSO also share similar coding principles (Grothe and Pecka, 2014). This has brought forward the notion that sound source position may not be estimated by MSO and LSO independently of each other, but rather that information from both nuclei is combined to retrieve one sound source estimate per hemisphere.

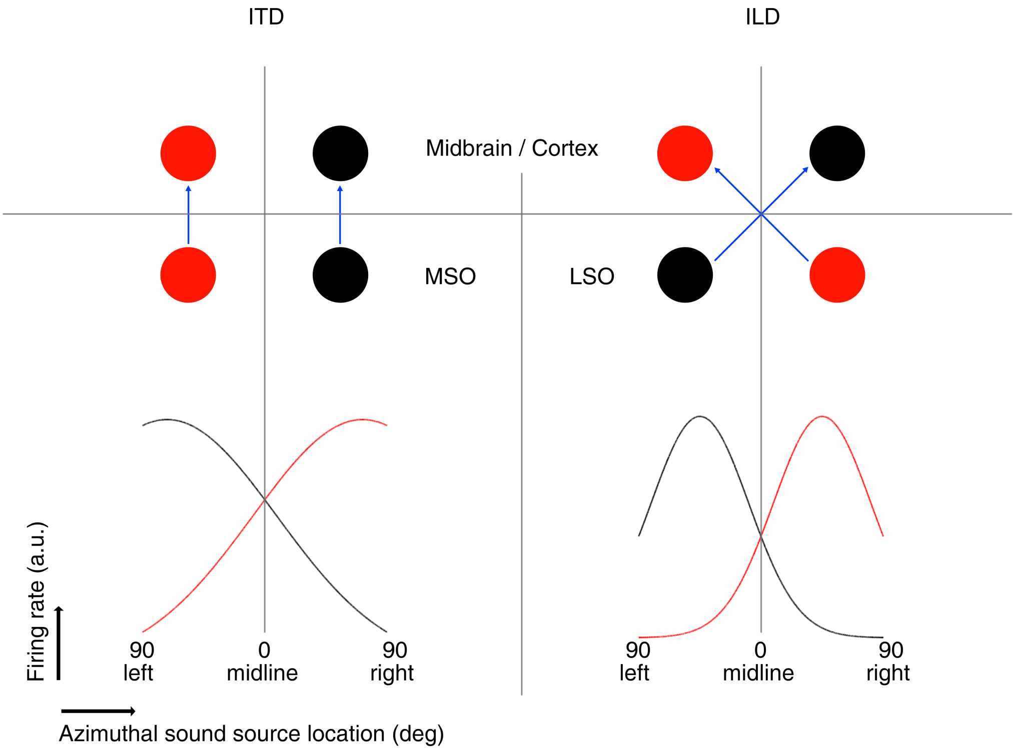

Figure 1.8: Hemispheric preference and tuning functions. Top graphs: Ipsilateral and contralateral MSO project to ipsilateral and contralateral midbrain, respectively. Ipsilateral and contralateral LSO project to contralateral and ipsilateral midbrain, respectively. Bottom graphs: Typical tuning curves (firing rate vs. sound source position) of MSO and LSO. Both tuning curves cover a large range of azimuthal space suggesting a 2-channel model of ITD/ILD coding. On Midbrain/Cortex level, similar firing rates in each channel would code for a sound positioned at 0° whereas higher firing rates in one hemisphere would code for a position more contralateral in comparison to the less active hemisphere. Overall activity (firing rate) is higher for sounds located contralaterally for MSO, for LSO is is higher for sounds located ipsilaterally. The cross-projections (top graphs) result in equal hemispheric preference for MSO and LSO as MSO maintains contralateral preference and LSO flips from ipsi- to contralateral preference. Figure inspired by Grothe and Pecka (2014).

Furthermore, typical tuning curves of MSO and LSO cover a wide range of auditory space (see Figure 1.8, bottom graphs). This further supports the idea of a left-right comparison model, i.e., that there might be two separate coding channels, one in each hemisphere, that may be compared and evaluated at higher auditory stages. This has further lead to the concept of *hemispheric balancing* (Lingner et al., 2018) where each channel (or hemisphere) computes one estimation of the position of a sound and the two estimates are then combined with each other and balanced according to the responses of MSO and LSO, i.e., the higher the activity in one hemisphere is, the more the estimate of this hemisphere will weigh into the balanced estimate. The mathematical implementation of hemispheric balancing is introduced in Section 2.2.1.
Another important fact about the MSO and LSO tuning curves and their implications for *hemispheric preference* should be stressed at this point. The MSO responds best to sounds located contralaterally (see 1.8, left bottom) whereas the LSO responds best to sounds located ipsilaterally (see 1.8, right bottom). This does, however, not present a conflict when evaluating the responses at midbrain or cortex level in one hemisphere, because the MSO has projects to ipsilateral IC (excitatory) and the LSO projects to contralateral IC (also excitatory, see 1.8, top graphs; also cf. Section 1.1.4). Thus, whereas the MSO preserves its contralateral preference at higher auditory stages, the LSO flips its ipsilateral preference to a contralateral preference (Beckius et al., 1999; Grothe and Pecka, 2014). These are features that need to be incorporated to gain a biologically correct phenomenological description of the neuron populations of these two nuclei.

1.5 AIMS OF THESIS

The neuronal code underlying sound localization in azimuthal space is unknown. In this thesis, we present a novel model of binaural hearing and apply it to complex auditory scenes in order to separate sounds as a listener would do in a cocktail party setting. To achieve this, the thesis can be subdivided into three main aims.

1. The Effective Model.
   The first aim is to design a complete biologically moti-

vated computational model that captures large parts of the phenomenology of the human binaural hearing apparatus. We use this model to propose a novel neuronal code for auditory space.

2. Short-Term ITD Estimation.
   Secondly, we use the effective model and the neuronal code to estimate positions of stimuli in horizontal space. For this we make use of model-based ITD estimations in very short time bins, hence the name short-term ITD estimation.

3. Sound Separation.
   Finally, we use the estimated ITDs to separate sounds from each other. We show that it is in principle possible to perform sound separation in complex auditory scenes solely based on biologically processed ITDs.

2 RESULTS

All models are wrong, ...

— George E. P. Box

2.1 MODEL ANATOMY

2.1.1 *Effective Models of MSO and lLSO*

In order to simulate the processing of sounds in the two ITD-sensitive nuclei MSO and lLSO in a fast manner, we implement a structure that we coin an *effective* model for each of these two nuclei. It is termed an effective model because we simulate the individual neurons according to three defining properties which results in a full characterization of each neuron. These three properties are the best frequency (BF), the characteristic delay (CD) and the characteristic phase (CP). The BF is the frequency to which a neuron has its highest response, i.e., the highest firing rate as compared to neighboring frequencies. The CD is the frequency-independent time delay difference, i.e., the difference in time it takes for a sound to arrive from both ears at that neuron which is based on (morphological) properties that do not take sound frequency into account (e.g., axonal length). The CP is the frequency-dependent phase delay difference, i.e., the difference in phase it takes for a sound to arrive from both ears at that neuron which is based on properties that take sound frequency into account (e.g., cochlear filtering). These three properties – BF, CD and CP – are connected via the best ITD

$$\mathrm{ITD}_{\mathrm{best}} = \mathrm{CD} + \frac{\mathrm{CP}}{f} \tag{2.1}$$

which is the interaural time difference where a neuron has its highest firing rate (Yin and Kuwada, 1983b; Yin and Chan,

1990; Pecka et al., 2008). The denominator f in (2.1) refers to the frequency of the current input sound. CD and CP are fixed values for each neuron. Note that f is given in Hz, CD in seconds and CP in cycles in order to obtain ITD_{best} in seconds.
The term *effective* in our model therefore stresses the fact that we can bypass the explicit modeling of each step of excitation and inhibition of the ascending auditory pathway as seen in Figure 1.3 and 1.4, because the result of these processes is already captured by these three defining properties of each ITD-sensitive neuron. The distribution of these parameters is discussed in the following subsection.

2.1.2 *MSO and lLSO Populations*

To achieve a realistic model of the MSO and lLSO in humans, we have adjusted the total number of neurons in each nucleus and the characteristics BF, CD, and CP to fit recent data from Hilbig et al. (2009) and Lüling et al. (2011).

Population Sizes
According to Hilbig et al. (2009), Nissl staining in human auditory brainstem nuclei and explicit counting of neurons via Kontron Image Analysis (Zeiss, Germany), approximated the MSO of humans to be comprised of $3,891$ and the LSO of $1,980$ neurons. Considering computational power limitations of the model, we set the total number of MSO neurons to 100 (per BF). The number of lLSO neurons (per BF) is set to 51 due to the ratio LSO/MSO. The exact size of the lLSO in humans is unknown. In Section 3 (Discussion), we discuss varying lLSO population sizes and their impact on the ITD encoding mechanism as described in Section 2.1.4. For left and right hemisphere, we model the same amount of neurons with identical distribution of the characteristics CP and CD. Because we analyze eight different BFs per hemisphere, the total sum of neurons for both hemispheres in the model is therefore 1600 for the MSO and 816 for the lLSO.

Distribution of BFs
The distribution of BFs is based on the physiological properties of the basilar membrane as discussed in Section 1.1.3. The apex

frequency in our model is set to 200Hz, the highest frequency is set to 1500Hz to ensure that we stay in ITD-relevant frequency range. We then use logarithmic spacing to create eight different BFs. The formula for the n-th BF is given by

$$\mathrm{BF}(n) = f_0 \times \left(\frac{f_N}{f_0}\right)^{n/7} \tag{2.2}$$

where $f_0 = 200$, $f_N = 1500$ and $n \in \{0, 1, ..., 7\}$. For the exact filter implementation, which introduces the BF, see Section 1.1.3 (Basilar Membrane).

Distribution of CPs and CDs

The choice for the distribution of CPs and CDs is tuned in such a way to reflect the negative correlation between CP and CD as reported by Lüling et al. (2011). The data in this study is based on recordings from DNLL neurons in gerbils (*Meriones unguiculatus*). However, the DNLL is the subsequent downstream nucleus after the SOC and receives input from MSO and LSO neurons. Furthermore, the DNLL also preserves the ITD sensitivities from these nuclei (Siveke et al., 2006). Therefore it is a straightforward choice to assume that the distribution of CPs and CDs from MSO and lLSO can be equated to the distributions as described in the DNLL. For the MSO, we assume uniformly distributed CPs ranging from $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$ radians. For the lLSO, which exhibits larger CPs, we assume them to be uniformly distributed ranging from $+\frac{\pi}{2}$ to $+\frac{3\pi}{2}$ radians. CP values are motivated by Lingner et al., 2018. The numbers are created in MATLAB (The MathWorks, Inc., Natick, Massachusetts, US) using a Mersenne Twister pseudorandom number generator with a fixed seed of 0 to ensure reproducibility. The CDs are derived from the CPs via equation (2.1). Solving for CD for every neuron for each fixed BF results in

$$\mathrm{CD} = \frac{\mathrm{ITD}_{\mathrm{best}} \times \mathrm{BF} - \mathrm{CP}}{\mathrm{BF}} = \frac{\mathrm{BP} - \mathrm{CP}}{\mathrm{BF}} \tag{2.3}$$

where the last equality is due to the relationship $\mathrm{ITD}_{\mathrm{best}} \times \mathrm{BF} = \mathrm{BP}$ which is referred to as the *best phase* (BP). To obtain CD in cycles, the values for CP are transformed from radians into cycles before entering equation (2.3) via $\frac{\mathrm{CP}}{2\pi}$, BP is given in cycles and f in Hz. For the MSO we assume the BPs to be normally distributed with mean $\mu = 0.085$ cycles and standard deviation $\sigma = 0.05$ cycles. For the lLSO we assume the BPs to be normally distributed with mean $\mu = 0.325$ cycles and standard deviation $\sigma = 0.05$ cycles. BP values are motivated by Lingner et al., 2018. The numbers are generated with the MATLAB in-built normal random generator $\mathtt{normrnd}$ (with a fixed seed of 0 for reproducibility). For our model, we assume these parameter distributions to also hold true for humans (see Section 3 (Discussion)). Note that the choice of parameters guarantees the negative correlation as seen in Lüling et al. (2011).

Summary of Population Set-Up

The settings of all model parameters of one hemisphere are summarized in the following Table 2.1. We assume identical parameter distributions for each of the two nuclei in each hemisphere, but for every BF there is a different set of CDs and CPs chosen. In one hemisphere, there are thus no two identical CDs or CPs, respectively, i.e., no two neurons of the same hemisphere have the same distribution of BF, CD or CP; guaranteeing uniqueness of every neuron.

	MSO	lLSO
# of neurons	800	408
BFs [Hz]	$[200, 1500]$	$[200, 1500]$
BPs [cycles]	$\mathcal{N}(0.085, 0.05)$	$\mathcal{N}(0.325, 0.05)$
CPs [radians]	$\mathcal{U}(-\frac{\pi}{2}, +\frac{\pi}{2})$	$\mathcal{U}(+\frac{\pi}{2}, +\frac{3\pi}{2})$

Table 2.1: Population parameter distributions in each hemisphere. The eight BFs are logarithmically scaled in the given range. $\mathcal{N}(\mu, \sigma)$ denotes the normal distribution with mean μ and standard deviation σ. $\mathcal{U}(\alpha, \beta)$ denotes the uniform distribution in the range α, β.

The from Table 2.1 resulting overall MSO and lLSO populations are shown in Figure 2.1. Note that the y-axis is frequency-scaled as CD x BF. This is done to ensure that the here modeled populations demonstrate the aforementioned same negative correlation between CD x BF and CP (steepness of slope roughly -1) as observed in Lüling et al. (2011). Furthermore, the distributions of the parameters directly reflect that the lLSO projects contralaterally to IC and the MSO ipsilaterally (cf. Chapter 1 (Introduction)). This is due to the fact that the larger CPs of the lLSO, as compared to the MSO, result in negative CDs as a consequence of equation (2.3).

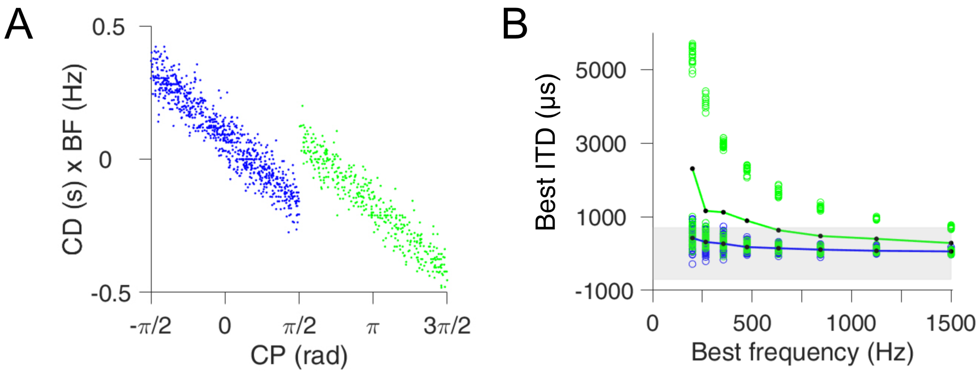

Figure 2.1: Neuronal properties. A. Distribution of the population of neurons for MSO and lLSO in each hemisphere using the frequency-scaled parameter CD x BF vs. CP. MSO population in blue, lLSO in green. Note that the lLSO has larger CPs than the MSO, resulting in CD x BF < 0 which indicates the aforementioned contralateral projection to the IC (see Section 1.4.1). B. Distribution of best ITD as defined in equation (2.1) vs. best frequency for all MSO (blue) and lLSO (green) neurons of the left hemisphere with respective mean best ITD curves (solid colored lines). Black dots on colored lines refer to the mean best ITD to a fixed BF-channel. Physiological range of humans denoted by the gray shaded area.

2.1.3 *Model Stages*

The model can in its entirety be described as an input-output function. The *input* signal is an arbitrary binaural pressure wave. Binaural, because for each signal we calculate the input at the left and right ear separately. The *output* of the model is the membrane potential of the MSO or lLSO in one hemisphere where the information from the left and right ear is combined. The model can be divided into three stages. The first stage is the introduction of the ITD. The second and third stages are the peripheral and the binaural processing, respectively. An overview is shown in Figure 2.2. The model is implemented in MATLAB.

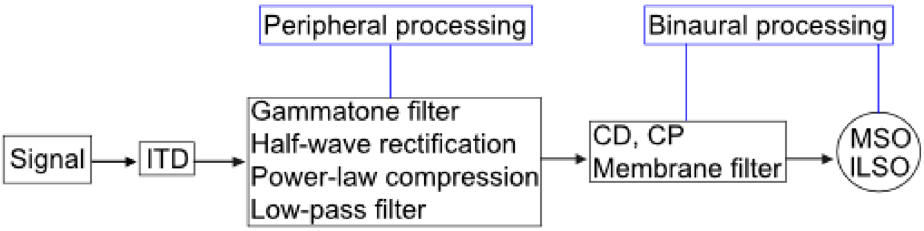

Figure 2.2: Model flow diagram. An ITD is added to the input signal which then goes through peripheral and binaural processing, resulting in a membrane potential at the MSO or lLSO.

Signals

There are two types of input signals. The first type is a 1 second long sinusoidal pure tone $s(t) = sin(2\pi Ft)$ ($t \geqslant 0$ time in seconds) with the frequency F in hertz set to the current frequency of the observed BF (in order to maximally drive the neurons). The second type is a complex signal which is a 1 minute long excerpt from the audiobook *Gone Girl* (© Random House Audio) with a US-American male and US-American female speaking in alternation, both of equal length.
To ensure that all model input signals, be it pure or complex, have the same loudness, all inputs are normalized to correspond to $60dB\ SPL$ (sound pressure level). This value has been chosen by us as a base reference value to be able to scale different sounds with different SPLs and make them comparable (see Section 3.2.3). The sinusoidal tones are generated directly in MATLAB. The complex signal has been recorded directly from the audiobook using Audacity (Audacity® is a registered trademark of Dominic Mazzoni). Speech pauses in the complex signal have been cut out to get a maximum of $0.5s$ pauses between sentences. All signals were sampled at $96kHz$. The choice for $96kHz$ rather than the standard $44.1kHz$ is owing to the better resolution of the time axis (roughly 22.68µs for $44.1kHz$ vs. 10.42µs for $96kHz$). This gives better control and numerical stability for shifting signals along the time axis on a microsecond scale.

ITD Introduction

The ITD implementation refers to how the binaural pressure wave reaches the two separate ears depending on the position of the sound in the azimuthal plane. For our model we use the following convention: Positive ITDs refer to a sound coming from the right with respect to midline, meaning that a sound will also reach the right ear earlier than the left ear. Negative ITDs, in turn, refer to a sound coming from the left. An ITD of zero refers to a sound coming directly from straight ahead (0° azimuth/midline). The input in our model at the right and left ear are calculated and processed separately, therefore the first step is to have two identical copies of the input signal $s(t)$, i.e., $s_R(t)$ and $s_L(t)$ with R/L referring to the input at the right and left ear, respectively. For positive ITDs, we implement the ITD by shifting the onset of the input signal at the right ear by $\frac{\text{ITD}}{2}$ to the *left* on the time axis. Analogously, we shift the beginning of the sound at the left ear by the same summand to the *right*. This gives the generalized form $s_R(t + \frac{\text{ITD}}{2})$ and $s_L(t - \frac{\text{ITD}}{2})$ which results in the sound arriving at the right ear first with the onset of the sound at the left ear being delayed by exactly 2 x $\frac{\text{ITD}}{2} = \text{ITD}$. The leading ear, i.e., the ear where the sound arrives first, switches from right to left for $\text{ITD} < 0$. Following ITD introduction, the separate inputs at both ears $s_{R/L}$ go through peripheral processing as described in the next subsection.

Peripheral Processing

As decribed in Section 1.1.2, the auditory periphery encompasses those stages of processing that start at the pinna (i.e., the arrival of the pressure wave at the ear) and end at the auditory nerve (AN) where then action potentials transport the information to the MSO and lLSO in the auditory brainstem (Meddis and Lopez-Poveda, 2010). The model equivalent to the entire peripheral processing stage is implemented by our cochlear model which is biophysically motivated by how the inputs are (non-linearly) transformed at the basilar membrane (BM) and subsequently transduced at the inner hair cells (IHCs).

Pre-Processing. First of all, in order to dampen effects introduced by cochlear filtering at the BM, the two input signals from both ears have to be pre-processed. The first and last $10\mathrm{ms}$ of each signal are multiplied with a ramp function, i.e., $\tilde{s}_{R/L}(t) = s_{R/L}(t) \text{ x } R(t)$ and R is given by

$$R(t) = \begin{cases} +100t & ; \ t < \text{first 0.01s of signal } s_{R/L} \\ 1 & ; \ \text{else} \\ 1 - 100t & ; \ t > \text{last 0.01s of signal } s_{R/L} \end{cases}$$

where the factors ± 100 are given in Hz and time t is measured in seconds, rendering R dimensionless. The ramp R prevents extreme amplitude distortions of the input signal when the BM corresponding filter is applied.

Basilar Membrane. The BM consists of a gammatone filter bank which introduces the BF to the pre-processed signals. Each different BF in the model resembles a different point of stimulation along the BM from lower frequencies (from the apex of the BM) towards higher frequencies (towards the base of the BM). As we model eight different BFs, we thus have a bank consisting of eight different gammatone filters. In time domain, the impulse response function of the gammatone filter $\gamma(t)$ is given by

$$\gamma(t) = \alpha t^{(n-1)} \exp(-2\pi t\beta) \cos(2\pi t\phi) \tag{2.4}$$

where n is the order of the filter, α the amplitude, β the bandwidth and ϕ the filter's center frequency (Patterson and Moore, 1986). For simplicity, at this stage in the model we assume $\alpha = 1$ (owing to later introduced BF-dependent amplitude modulations as described in Section 2.1.4). A filter order of $n = 4$ has been found to be a good fit to describe the human auditory filter (Patterson et al., 1992). The bandwidth β gives the duration of the filter and is defined as

$$\beta = \mathrm{ERB}/b_n$$

where ERB is the equivalent rectangular bandwidth and b_n is an order-dependent factor given by

$$b_n = \frac{\pi(2n-2)!2^{-(2n-2)}}{[(n-1)!]^2}$$

In our case ($n = 4$) the value is $\frac{20\pi}{2^6} \approx 0.9817$. The ERB describes the varying bandwidths of the gammatone filter along the BM. For humans, the ERB is approximated by

$$\mathrm{ERB} = 24.7 + 0.108\phi \tag{2.5}$$

(Glasberg and Moore, 1990) and thus the ERB is dependent on the center frequency ϕ which is set to the current observed BF. Thus, this part in the model is where the first of our three neuron-defining properties (BF, CD and CP) is introduced. The equations (2.4) and (2.5) fully characterize the gammatone filter bank.
The two pre-processed input signals are convolved with the gammatone filter, i.e.,

$$\widehat{s}_{R/L}(t) = (\widetilde{s}_{R/L} * \gamma)(t)$$

and the response is passed on to the IHCs.

Inner Hair Cells. The response of the basilar membrane is now transduced by the IHCs into an electric potential $p(t)$ which models the release of transmitters onto spiral ganglion cell dendrites. This signal transduction consists of four steps. First, the output of the BM is half-wave rectified, i.e., $\widehat{s}_{R/L}(t_0) = 0$ if $\widehat{s}_{R/L}(t_0) < 0$ for all times t_0. The result is then power-law compressed to the power of 0.4 (Oxenham and Moore, 1994), i.e., $c_{R/L}(t) = |\widehat{s}_{R/L}(t)|^{0.4}$. The compressed signal c is then convolved with a low-pass filter l of second order and frequency cut-off at 1kHz (Lingner et al., 2012), i.e., $p_{R/L}(t) = (c_{R/L} * l)(t)$. The output of p of the cochlear model is the result of the peripheral processing and is now passed on to the binaural processing stages. Finally, to also account for the latency that occurs during mechano-electrical transduction, we introduce a uniformly distributed neuron-fixed latency jitter from the range of $]0;1[$ milliseconds (Rhode and Smith, 1986; Ford et al., 2015). The numbers are generated – analogously to the CPs – with a Mersenne Twister pseudorandom number generator and a fixed seed of 0 to ensure reproducibility. The latency jitter is implemented via a time-shift, i.e., $j_{R/L}(t) = p_{R/L}(t - \text{Jitter})$ and passed on to the stage of binaural processing.

Binaural Processing

The term binaural processing in the ascending auditory pathway in our model refers to the first time where neurons in the MSO and lLSO use information from both ears. That information is then combined to generate membrane potentials in these two nuclei. As described in Section 2.1, we implement an effective model to bypass modeling of the exact stages of inhibition and excitation as shown in Section 1.1.4. The binaural information in our model is captured in its entirety by the two neuron-defining properties of CD and CP.

Characteristic Delays. The CD is the difference in arrival time it takes for a sound to arrive at one neuron from both ears and can thus be implemented by a simple time shift. Similarly to the ITD, this is achieved by shifting the response $j_{R/L}$ by $\frac{CD}{2}$ into the corresponding direction on the time axis, i.e.,

$$r_R(t) = j_R(t - \frac{CD}{2}) \text{ and } r_L(t) = j_L(t + \frac{CD}{2})$$

The sign is determined by equation (2.1) and our convention that sounds arrive at the right ear first when ITDs are greater than zero.

Characteristic Phases. The CP is the difference in phase it takes for a sound to arrive at one neuron from both ears and can thus not be implemented by a simple time shift but rather by a phase shift. However, the choice and reasoning for the sign of the shift is the same as for the CD, i.e., r_R is shifted by $-\frac{CP}{2}$ and r_L is shifted by $+\frac{CP}{2}$.
In order to find an appropriate phase-shifting expression that can easily be applied to non-trivial sounds, we look at the functions $r_{R/L}(t)$ in frequency-domain. For reasons of comprehensibility, let us first look at $r_R(t)$. The function of frequency, or the Fourier transform, of $r_R(t)$ is given by

$$\widehat{r}_R(\omega) = \int_{-\infty}^{\infty} dt e^{-i\omega t} r_R(t) \tag{2.6}$$

where ω represents the angular frequency and i is the imaginary unit defined as $i^2 = -1$. To shift the Fourier transform $\widehat{r}_R(\omega)$ by phase Φ, we simply multiply $\widehat{r}_R$ with $\exp(\Phi \text{sgn}(\omega) i)$ where $\text{sgn}(\omega)$ denotes the sign-function given by

$$\text{sgn}(\omega) = \begin{cases} 1 & ; \ \omega > 0 \\ 0 & ; \ \omega = 0 \\ -1 & ; \ \omega < 0 \end{cases}$$

For $\Phi = -\frac{CP}{2}$ we get

$$\widehat{q}_R(\omega) = \widehat{r}_R(\omega) \exp\left(-\frac{CP}{2} \text{sgn}(\omega) i\right) \tag{2.7}$$

as the phase-shifted version of $r_R(\omega)$. The Fourier Convolution Theorem (FCT) states that the back transform of the product of two Fourier transforms is the convolution of the back transforms of the two factors in time domain, i.e., for two functions f_1 and f_2 the FCT can be written as

$$\widehat{f}_1(\omega)\widehat{f}_2(\omega) \quad \leftrightarrow \quad (f_1 * f_2)(t) \tag{2.8}$$

The back transform of $\widehat{q}_R(\omega)$ is therefore – according to the FCT (2.8) – given by

$$q_R(t) = (r_R * u_R)(t) \tag{2.9}$$

where $u_R(t)$ denotes the back transform of $\widehat{u}_R(\omega) = e^{-\frac{CP}{2}\mathrm{sgn}(\omega)i}$ and $r_R(t)$ is already known. Hence, finding the back transform of $\widehat{u}_R(\omega)$ solves the problem of shifting r_R by phase $-\frac{CP}{2}$ by performing the convolution described in (2.9).
In time-domain, the back transform $u_R(t)$ of $\widehat{u}_R(\omega)$ is defined as

$$u_R(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega e^{+i\omega t}\widehat{u}_R(\omega) \tag{2.10}$$

The improper integral (2.10) does not converge in the time domain for this specific choice of $\widehat{u}_R(\omega)$. To establish an improper integral that is finite and exists, we multiply (2.10) with a *regularization factor* given by

$$e^{-\epsilon|\omega|} \tag{2.11}$$

where $|\cdot|$ denotes the absolute-value-function or modulus defined as

$$|x| = \begin{cases} x & ; \ x \geqslant 0 \\ -x & ; \ x < 0 \end{cases}$$

for $x \in \mathbb{R}$. The reason for the choice of the regularization factor (2.11) is because it turns into a multiplicative constant of 1 for infinitesimal small $\epsilon > 0$, because

$$\lim_{\epsilon \to 0} e^{-\epsilon|\omega|} = e^0 = 1$$

regardless of the value of ω. To this end, for very small $\epsilon > 0$, we rewrite (2.10) as

$$\begin{aligned} u_R(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega e^{+i\omega t}\widehat{u}_R(\omega)e^{-\epsilon|\omega|} \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega e^{+i\omega t}e^{-\frac{CP}{2}\mathrm{sgn}(\omega)i}e^{-\epsilon|\omega|} \end{aligned} \tag{2.12}$$

Splitting the integral into its negative and positive parts gives us

$$\frac{1}{2\pi}\left(\int_{-\infty}^{0} d\omega e^{+i\omega t}e^{+\frac{CP}{2}i}e^{+\epsilon\omega} + \int_{0}^{\infty} d\omega e^{+i\omega t}e^{-\frac{CP}{2}i}e^{-\epsilon\omega}\right)$$
$$= \frac{1}{2\pi}\left(e^{+\frac{CP}{2}i}\int_{-\infty}^{0} d\omega e^{\omega(it+\epsilon)} + e^{-\frac{CP}{2}i}\int_{0}^{\infty} d\omega e^{\omega(it-\epsilon)}\right)$$

Calculating the anti-derivatives of the summands yields

$$\frac{1}{2\pi}\left(e^{+\frac{CP}{2}i}\left[\frac{1}{it+\epsilon}e^{\omega(it+\epsilon)}\right]_{-\infty}^{0} + e^{-\frac{CP}{2}i}\left[\frac{1}{it-\epsilon}e^{\omega(it-\epsilon)}\right]_{0}^{\infty}\right) \tag{2.13}$$

The vanishing of the non-trivial limit

$$\lim_{\omega\to\infty} e^{\omega(it-\epsilon)} \tag{2.14}$$

can be seen as follows. We rewrite (2.14) as

$$\lim_{\omega\to\infty}\left(e^{\omega i}\right)^{t}\lim_{\omega\to\infty} e^{-\epsilon\omega} \tag{2.15}$$

The expression

$$e^{\omega i} = \cos(\omega) + i\sin(\omega) \tag{2.16}$$

is known as *Euler's formula* and describes that $e^{\omega i}$ is a finite (complex) number located on the circle with origin at 0 and a radius of 1. This means that $\left(e^{\omega i}\right)^{t}$ is an arbitrary oscillatory factor because $t > 0$ is also finite. The limit in (2.15) therefore only depends on $\lim_{\omega\to\infty} e^{-\epsilon\omega}$ which is 0 for $\epsilon > 0$. Consequentially, the limit (2.14) also vanishes. Analog reasoning gives

$$\lim_{\omega\to-\infty} e^{\omega(it+\epsilon)} = 0 \tag{2.17}$$

Plugging these limits into equation (2.13) and using $\exp(0) = 1$ yields

$$\begin{aligned}&\frac{1}{2\pi}\left(e^{+\frac{CP}{2}i}\left(\frac{1}{it+\epsilon}\right) + e^{-\frac{CP}{2}i}\left(-\frac{1}{it-\epsilon}\right)\right)\\ &= \frac{1}{2\pi}\left(\frac{e^{+\frac{CP}{2}i}(-it+\epsilon) + e^{-\frac{CP}{2}i}(it+\epsilon)}{\epsilon^2+t^2}\right)\end{aligned} \tag{2.18}$$

and applying Euler's formula (2.16) for $\omega = \pm\frac{CP}{2}$ reduces the equation to

$$\frac{(\cos(\frac{CP}{2}) + i\sin(\frac{CP}{2}))(-it + \epsilon) + (\cos(\frac{CP}{2}) - i\sin(\frac{CP}{2}))(it + \epsilon)}{2\pi(\epsilon^2 + t^2)} \tag{2.19}$$

because $\sin(-x) = -\sin(x)$ and $\cos(-x) = \cos(x)$ for $x \in \mathbb{R}$. Finally, simplifying (2.19) gives us the convolution kernel

$$u_R(t) = \frac{1}{\pi(\epsilon^2 + t^2)}\left(t\sin\left(\frac{CP}{2}\right) + \epsilon\cos\left(\frac{CP}{2}\right)\right) \tag{2.20}$$

and $q_R(t) = (r_R * u_R)(t)$ performs the phase shift in time-domain. Replacing $\frac{CP}{2}$ with $-\frac{CP}{2}$ gives us the kernel for $r_L(t)$, i.e.,

$$u_R(t) = \frac{1}{\pi(\epsilon^2 + t^2)}\left(-t\sin\left(\frac{CP}{2}\right) + \epsilon\cos\left(\frac{CP}{2}\right)\right)$$

and $q_L(t) = (r_L * u_L)(t)$ implements the phase shift.
In our model, setting $\epsilon = \frac{20}{96000} \approx 0$ for the regularization factor results in robust phase shifts.

Membrane Potentials. To translate the response of the binaural processing steps into a membrane potential, we must account for the biophysics of typical MSO and lLSO membranes as described in Section 3.3.1. Therefore, we designed a filter to reflect the channel kinetics, i.e., the interplay of the low-threshold voltage-gated potassium channels (Kv1), the hyperpolarization-activated cyclic nucleotide-gated cation channels (HCN) as well as account for the passive membrane properties with its low-pass frequency behavior due to the leakiness of MSO and lLSO neurons. Following the notions in Fischer et al. (2018), we use a second-order gammatone-bandpass-filter of the form

$$m(t) = t\exp(-t\beta_m)\cos(2\pi t\Phi_m) \tag{2.21}$$

which acts as membrane filter m which is applied to the CD- and CP-shifted responses $q_{R/L}(t)$. The bandwidth is set to $\beta_m = 2765\text{Hz}$ and the center frequency to $\Phi_m = 427\text{Hz}$. This type of filter cascades a low-pass and high-pass filter and, most importantly, the filter parameter choices establish a membrane time constant of $\tau_m = \frac{1}{\beta_m} = \frac{1}{2765\text{Hz}} \approx 361.66\mu\text{s}$ which is small enough to guarantee the existence of membrane potential resonance properties which facilitate frequency specific processing (Fischer et al., 2018). Convolved with m, i.e., $V_{R/L}(t) = (q_{R/L} * m)(t)$ we receive the individual inputs from right and left ear to the MSO or lLSO cells. The membrane potential of one cell is the sum of these inputs, i.e.,

$$V(t) = V_R(t) + V_L(t) \tag{2.22}$$

and the resulting output, i.e., the amplitude, is interpreted as the membrane potential $V(t)$ in response to a specific sound for fixed parameters BF, CD and CP.

2.1.4 *Population ITD Encoding*

Neuronal Subpopulations. For every single cell in the MSO and lLSO we derive a separate membrane potential $V(t)$ as described in Equation 2.22. As mentioned in Section 1.2.2, a tone delay curve, i.e., the firing rate response of a cell vs. ITD, shows variability for pure tone sound stimuli with differing frequencies. Consequently, tone delay curves are not frequency-invariant. For this reason, we analyze subpopulations of neurons in each nucleus and hemisphere rather than interpreting the responses on a single-cell level. All cells that have the same BF assigned make up such a subpopulation which we now refer to as a *BF-channel* in order to emphasize the identical BF per neuron. Note, that this automatically implies that there is no crosstalk explicitly implemented between different BF-channels. The neurons are separated by frequency-channels. But due to the nature of the basilar gammatone filter as described in Section 2.1.3, this does not mean that there is an absolute separation of frequencies. There of course will be leak from higher frequency-channels into lower frequency-channels.

Tone Delay Curves Calibration. Frequency-variance is not the only problem at hand when analyzing auditory brainstem neurons. Several filtering mechanisms in the peripheral and binaural processing stages lead to inevitable distortions of the membrane potential $V(t)$. In order to correct for this, i.e., to extract tone delay curves that match the in-vivo data as observed in e.g., Yin and Kuwada (1983a) or Pecka et al. (2008), and also described in Section 1.2.2, the tone delay curves must first be calibrated via sinusoidal pure tones before being able to make implications about complex sounds.

The responses of all single neurons belonging to one BF-channel in each separate nucleus (MSO/lLSO) and hemisphere (left/ right) are summed up and averaged according to the number of neurons (100 for MSO and 51 for lLSO per BF-channel). This results in the population membrane potential. The driving stimulus for calibration is a sinusoidal pure tone with a frequency of the current observed BF. From this we plot the tone delay curve: population membrane potential against ITD. Each BF-channel then receives a membrane potential amplification factor A_{BF} in such a way that the peaks and troughs of the MSO and lLSO

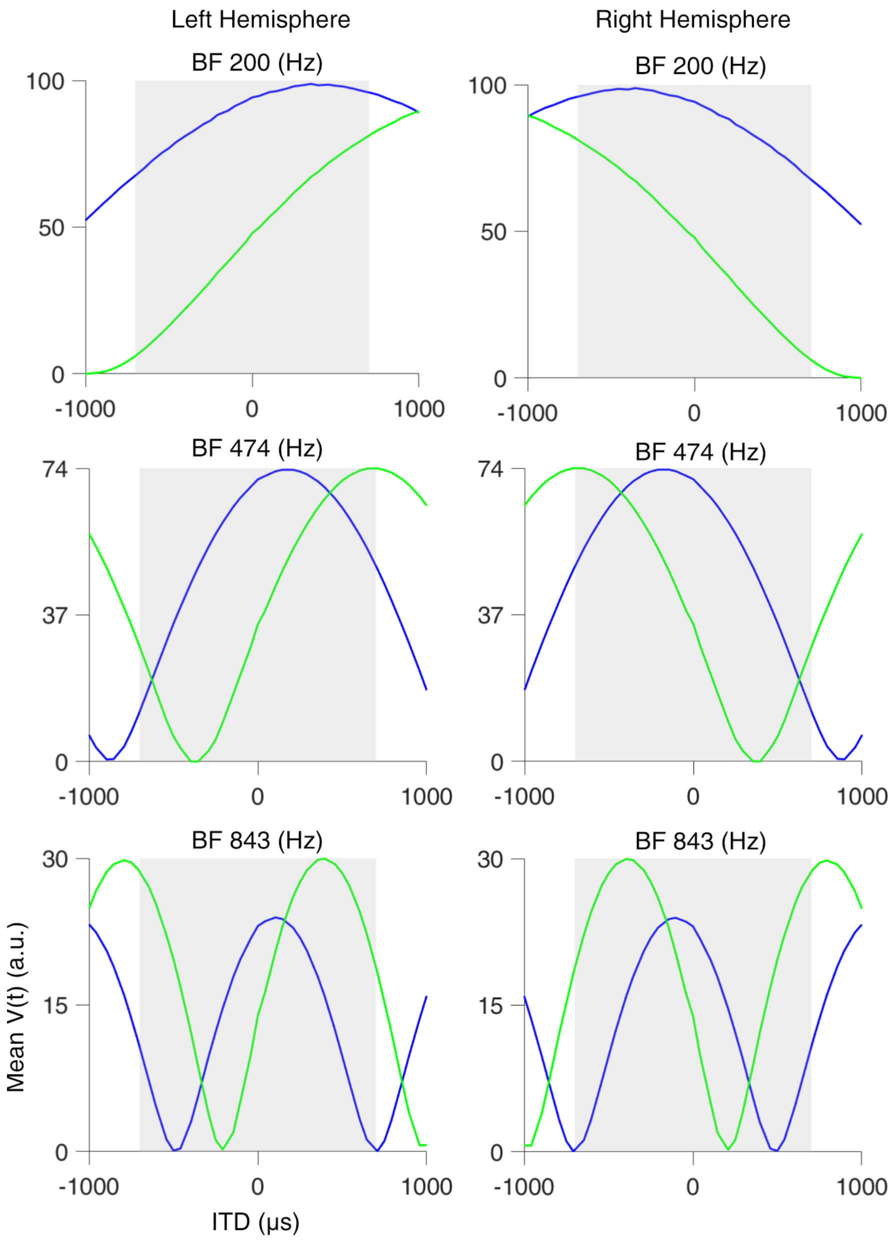

Figure 2.3: Tone Delay Curves. Mean $V(t)$ vs. ITD. Stimulus is a pure tone with frequency identical to BF. Each row corresponds to one specific BF-channel. MSO is depicted in blue, lLSO in green. Left panels show the response of the left hemisphere, right panels the response of the right hemisphere. Gray areas denote the physiological range for human (±700μs). The maximal average response of each nucleus declines the larger the value of the BF. Note how, for example, in the top left panel the MSO curve loses injectivity at the peak located at approximately +300μs. An ITD very close to the left or right to this point of injectivity-loss could not be resolved unambiguously by the MSO since one value of mean $V(t)$ would code for two different ITDs. However, the slopes of MSO and lLSO curves are offset (by approximately 500μs) and so the lLSO curve is still injective in an interval around +300μs meaning the lLSO curve can still resolve ITDs unambiguously as one mean $V(t)$ codes for exactly one ITD.

have similar population membrane potential. This is not ideally possible for all BF-channels as the MSO is more responsive to very low frequencies and loses range in the membrane potential domain the higher the value of BF is set (see bottom graphs in Figure 2.3). The lLSO also loses range in its membrane potential domain but is more stable than the MSO (which is to be expected for lLSO neurons). Once a fitting amplification factor A_{BF} per BF-channel is ascertained, this same factor is also applied to the membrane potentials which are generated from complex sound stimuli.

The resulting tone delay curves from three different BFs for sinusoidal stimuli can be seen in Figure 2.3. Note that the y-axis (population membrane potential) is in arbitrary units. The reason for this is that the exact absolute magnitude of response is not of importance for the model. It is rather crucial that the slopes of MSO and lLSO tone-delay curves are *offset*. Because this means, that if one of the two tone-delay curves (MSO or lLSO) loses injectivity due to reaching a peak or trough, the other delay curve can possibly counter this by being itself injective in an area around a current observed ITD and thus resolve the ITD unambiguously (see Figure 2.3 for further explanation).

Mean Population Response. To find an encoding mechanism for ITDs, we now look at the BF-channel-specific amplified responses. For this, we analyze these subpopulation responses under the condition that ITDs are fixed and plot the activity of the two nuclei of one hemisphere against each other, i.e., the activity from each hemisphere is evaluated *separately* at this point. The response lLSO vs. MSO is then plotted over time t (in seconds) for the entire duration of the input stimulus. In Figure 2.4 we see the 200Hz-channel population response from the left hemisphere for a sinusoidal driven at 200Hz (top graphs) and for the complex calibration sound stimulus (bottom graphs). For the right hemisphere, the resulting graphs have the same shape due to the hemispheric-symmetric structure of the model, but with the fixed ITDs having opposite sign.

For every fixed ITD (different ITDs are represented by different colors), every point on the respective colored curve is a sample with a sampling frequency of 96000Hz within a cycle of a sound. As the training stimuli are sufficiently long enough and span several cycles, this gives rise to ellipse-shaped geometric objects. Let us now consider all data points (i.e., the response to each point in time t) to one fixed ITD. We can interpret every data

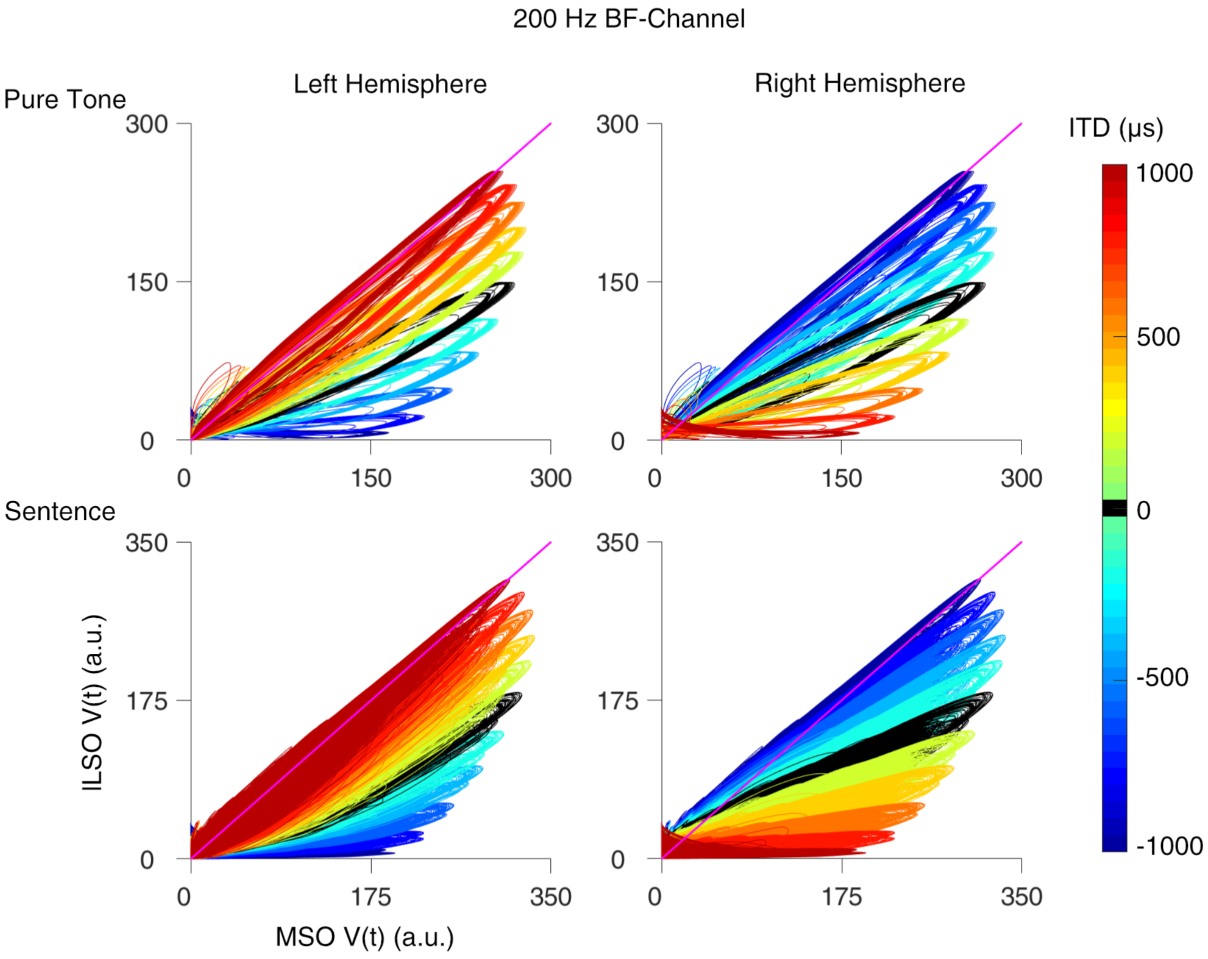

Figure 2.4: Mean population responses of lLSO vs. MSO at $\mathrm{BF} = 200\mathrm{Hz}$ for pure tone (top graphs) and sentence (bottom graphs). Different colors represent fixed ITD values. Identity line plotted in magenta for reference. Every population response vector $\vec{z}_{\mathrm{ITD}}$ encodes exactly one MPA $\phi_{\mathrm{ITD}} = \arg(\vec{z}_{\mathrm{ITD}})$ (see text for further explanation). For the left hemisphere (left graphs), ϕ_{ITD} is largest for high values of ITD. For the right hemisphere (right graphs), ϕ_{ITD} is largest for low values of ITD. The ellipses are denser for complex stimuli (bottom graphs), nevertheless the summation over all data points t results in a unique MPA for each individual ITD, establishing an encoding mechanism for ITDs.

point t as a complex number $z_t = a + ib \in \mathbb{C}$ ($a, b \in \mathbb{R}$ and i imaginary unit) in the Gaussian number plane. Summing up over all data points t

$$\sum_t z_t = \vec{z}_{\mathrm{ITD}} \tag{2.23}$$

results in the population response vector $\vec{z}_{\mathrm{ITD}}$ to a given fixed ITD. We now make use of a simple mathematical concept of complex numbers: Summing up over several complex numbers, which individually themselves can be represented by vectors in 2d-space, results in a single line or single vector. A simple consequence of vector algebra. For a perfect ellipse, this line would overlap with the principal axis of that ellipse. For the ellipse-similar objects as shown in Figure 2.4, the line can only be thought of as an approximation of such a principal axis. Nevertheless, summing up over a large set of data points t thus results in a stable line, the population response vector.

For every fixed ITD, i.e., for every ellipse-shaped object, the following procedure can be reiterated: We calculate the population response vector $\vec{z}_{\mathrm{ITD}}$ and propose that the ITD is encoded in this response vector via the argument

$$\phi_{\mathrm{ITD}} = \arg(\vec{z}_{\mathrm{ITD}}) \tag{2.24}$$

which is geometrically the angle enclosed by the axis of abscissas and the vector $\vec{z}_{\mathrm{ITD}}$. We coin this the *mean population angle* (MPA) to a given ITD. Calculating every MPA to every ITD thus provides an encoder for a specific ITD to a given BF. To make this encoder more precise, before calculating the MPA, a BF-specific noise limit is set to remove activity which is too low which would distort the MPA.

It is noteworthy at this point, that although the ellipses to different fixed ITDs overlap to some extent, this does not infer ITD ambiguity. Because if summed up over all data points, the resulting mean population vectors do not overlap anymore and, of course, the longer the training stimulus is, i.e., the longer you average over time, the robuster this vector is.

A general and crucial observation in the mean population responses is that the contraleading nucleus will always show a higher response than the ipsileading nucleus. For increasing positive ITDs (arriving at the right ear first), the contraleading left hemisphere shows an increase in the response of both nuclei. This also explains the overall form of the ellipses and how they change when manipulating the ITD from lower to higher values.

The principal axis of each ellipse not only gets expanded (i.e., the area of the ellipse enlarges), but the ellipses also move in an upwards direction when the ITD is shifted towards more positive values, peaking in its form at $+1000\mu s$. The same observations hold true vice versa for shifting from positive to negative ITDs when considering the right hemisphere. It is exactly this change of form of the ellipses which qualifies the MPA as an ITD encoding mechanism. The overall resulting ITD encoders, the *calibration curves*, are discussed in the next subsection and also the arising issue of ITD-ambiguity.

ITD Encoding: Calibration Curves. We now commence calibrating the model by calculating all MPAs to all combinations of ITDs and BF-channels per hemisphere. Calculations have been performed for ITDs ranging from $-1000\mu s$ (which denotes left with respect to midline) to $+1000\mu s$ (right with respect to midline) in steps of $50\mu s$. Note that we consider ITDs exceeding the physiological range of humans by $\pm 300\mu s$. First of all, in order to avoid computational errors in MATLAB when time-shifting signals, we need to take the sampling rate into account and adjust the ITD increments. The rate is 96000Hz and because $5 \text{ x } 10^{-5} \notin 96000^{-1}\mathbb{N}$ ($50\mu s$ is not a multiple of the inverse sampling rate), the increments are adjusted using the MATLAB in-built $\mathtt{round}$ function according to

$$\text{ITD}_{new} = (2 \text{ x } \frac{1}{96000}) \text{ x } \mathtt{round}\left(\frac{\text{ITD}}{2 \text{ x } \frac{1}{96000}}\right)$$

and the twice arising factor 2 is necessary because we shift by $\frac{\text{ITD}}{2}$ (see Section 2.1.3: ITD introduction). This adjustment of the ITD increments guarantees that the resulting ITDs are multiples of the inverse sampling rate, establishing a reliable time-shifting method and thus enabling us to keep faithful to the temporal precision of auditory brainstem cells on these very small time scales.

Calibration curves are calculated for each hemisphere individually for a fixed BF-channel. To every possible ITD the according MPA is computed. Plotting MPA vs. ITD gives us the the ITD encoder, the calibration curve. The resulting curves are shown in Figure 2.5. Different colors now represent different BF-channels. The top two panels show the calibration for the pure tone stimulus (driven at that BF), the bottom two panels show the calibration for the complex training stimulus. Left two panels represent left hemisphere and right two panels the right hemisphere. Note that the pure tone calibrations (frequency $F = BF$) show expected periodicity for multiples of the period $\frac{1}{F}$, which leads to non-injectivity when monotony switches at troughs and peaks. For higher frequencies, this presents a problem because the non-monotonic areas become very small with respect to the physiological range and thus a unique relation

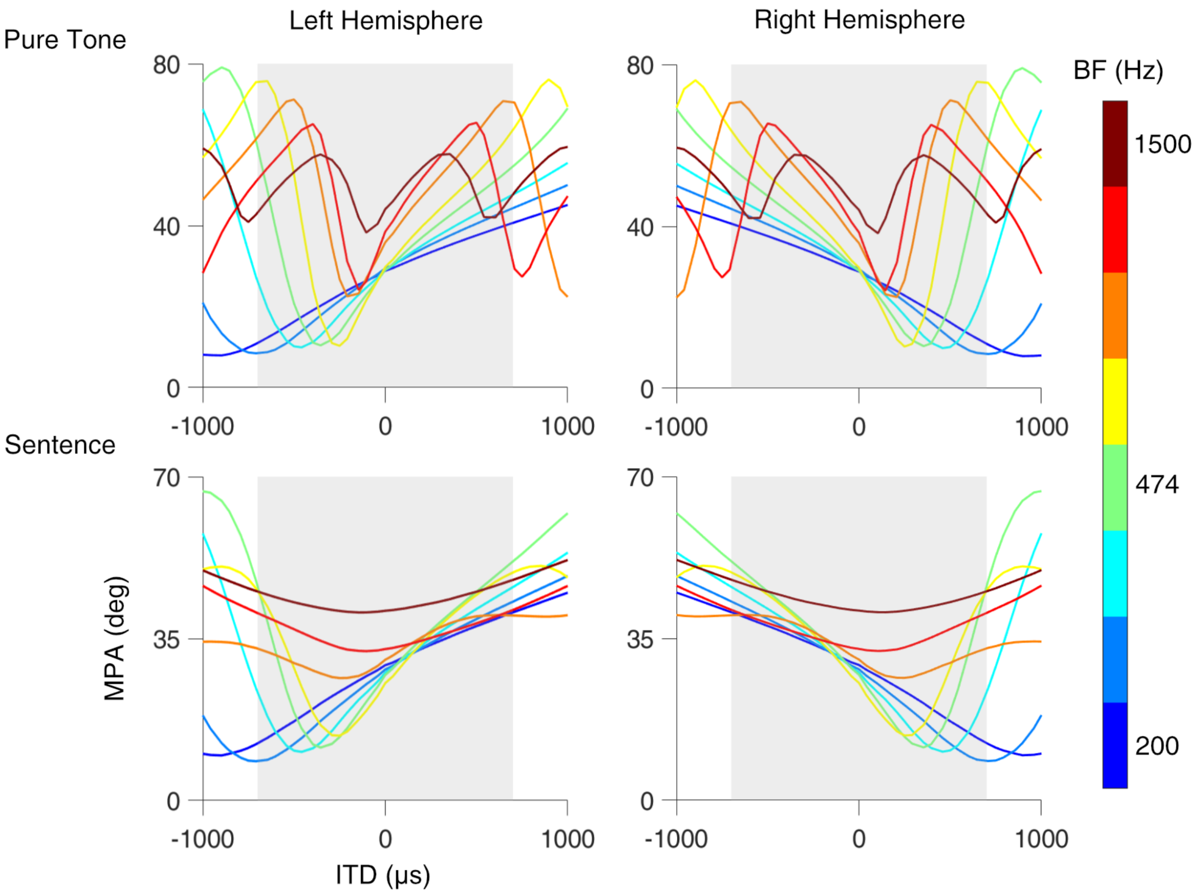

Figure 2.5: Calibration Curves. MPA vs. ITD for pure tone stimuli (top graphs) and complex stimuli (bottom graphs). Different colors represent different BF-channels. The calibration curves of the right hemisphere (right graphs) appear as mirrored (at axis ITD = 0) versions of the calibration curves of the left hemisphere (left graphs). ITDs are best encoded in areas of steep slopes and areas of curves that have similar MPA values represent frequency-invariant ITD encoding, i.e., regardless of the frequency of the current BF, the same ITD is encoded by very similar MPAs.

between MPA and ITD breaks down (Leibold and Grothe, 2015). For the complex sound stimulus, there is no ongoing periodic signal. The reason is because in sinusoidal tones the high frequency components are stationary and for spoken sentences these components are not stationary anymore, i.e., periodicity in the evaluated range is broken down. Simply put, this means that there is no clear periodicity visible in this range for complex

tones as it is for sinusoidal tones (compare Figure 2.5 top and bottom graphs). This is a direct consequence of the construction of the basilar membrane filter as described in Section 2.1.3. Due to the finite filter bandwidths (the ERB and thus also the filter width increases with increasing BF-channels), the IPDs (i.e., the interaural time differences multiplied with sound frequency) are no longer sufficiently constant for a given ITD at high frequencies. For complex tones which have a broad spectrum, this results in a wide range in which ITDs can later be decoded from as we have larger ITD-domains that are strictly monotonous. Although the calibration curves of the complex sound stimulus are not monotonous over the whole physiological range for BFs above 355.7Hz, the overall strictly monotonous ITD-ranges are still much larger than for periodic sounds. Thus the calibration curves of complex sounds provide a much better chance at decoding ITDs for arbitrary sound stimuli, especially for higher frequency channels.
Mechanistically speaking, the flattening of the curves in Figure 2.5 can be explained by the mean population responses for MSO and lLSO as shown in Figure 2.4. In that example, the responses are shown for $\mathrm{BF} = 200\mathrm{Hz}$. For higher frequencies (not shown) there are two important aspects that change. Firstly, the overall amplitudes becomes lower the higher the BF is. But since this occurs for both MSO and lLSO responses simultaneously, it would not influence the MPA since it is only a rescaling of both axes. Secondly, the different colored point clouds (ellipses which represent a fixed ITD) move closer together. The higher the BF, the more the ellipses move towards the midline axis, except those ellipses which are already clustered around midline as these stay stationary. A direct consequence from this phenomenon is, that the closer all of the ellipses are together, the more similar the MPAs become for varying ITDs, resulting in the calibration curves which can be observed in Figure 2.5.

ITD Decoding: Elimination Rule. To be able to use the calibration curves and successfully decode ITDs for arbitrary sound stimuli, ITD-ambiguity has to be resolved. To guarantee uniqueness of ITD, the calibration curve must be injective, i.e., one MPA may not encode more than one ITD. Geometrically this is only the case if a parallel to the ITD-axis has exactly one intersection with the calibration curve (Figure 2.5). Since this is not the case whenever a calibration curve has a trough or a peak, we have ITD-ambiguity for these cases. Restricting the calibration curves to monotonous parts counters the ITD-ambiguity directly. Since the left hemisphere encodes better for positive ITDs (contraleading) and the right hemisphere encodes better for negative ITDs, we implement the elimination rule as follows.

Elimination Rule: For left hemisphere, we only consider values of the calibration curves starting at the trough and the remaining *right* branches. For right hemisphere, we only consider the remaining *left* branches.

This straightforward rule can furthermore be justified by the fact, that those parts of the calibration curves which are eliminated via this rule yield very low mean membrane potentials $V(t)$ as these eliminated parts are located ipsilaterally with respect to the current hemisphere performing the ITD estimation. Thus ITD estimations in one hemisphere from those branches would not weigh much into the subsequent overall hemispherically balanced ITD estimation as described in Section 2.2.1, since it would be dominated by the ITD estimation from the other hemisphere due to the high mean membrane potential (contralateral preference). The elimination rule therefore can be seen as a means to make the hemispherically balanced ITD estimations more precise.

2.2 ITD estimations in complex auditory scenes and sound source separation

The model described in Chapter 2.1 can now be applied to estimate ITDs in complex auditory scences. Such a complex auditory scene is simulated by extending the number of inputs to the model. This can now be any number of n concurring, different signals (see Figure 2.6). Each signal has its own specific ITD assigned. All of the ITD-shifted signals are then summed up to obtain one superposed input signal. This signal now resembles a complex auditory scene with various sound sources distributed across the azimuthal plane. The effective model is then applied to this superposition input for each BF frequency-band and each hemisphere separately which results in an MSO and lLSO membrane potential. Similar to the derivation of the calibration curves, the membrane potentials are at first obtained over the whole stimulus duration time. However, estimating

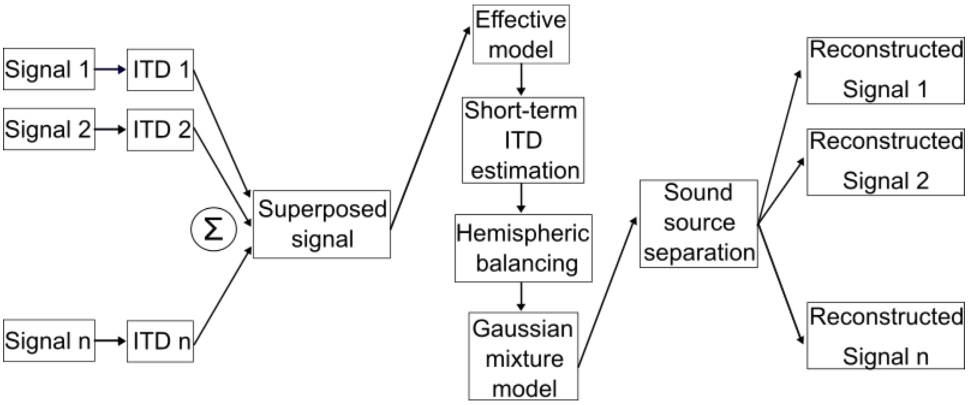

Figure 2.6: Complex auditory scenes and sound separation. Individual speakers are resembled by signals which each receive one ITD. The ITD corresponds to the signal's azimuthal position in space. The sum of all signals, the superposed signal, resembles a complex auditory scene. The effective model is applied to the superposed signal and subsequent short-term ITD estimation results in hemispheric estimates which are balanced according to each hemisphere's overall activity. A Gaussian mixture model determines the number of signals contained within the complex auditory scene and also their respective ITD. In a final step, the individual signals are extracted from the superposed signal (see main text for further explanation), yielding separated and reconstructed signals purely on the basis of ITD.

an ITD over the entire time interval (e.g., for a $2\mathrm{min}$ stimulus) would of course be biologically implausible. Instead, to retrieve biologically relevant results, we now introduce the concept of short-term ITD estimations.

2.2.1 *Short-Term ITD Estimations*

In general, to estimate ITDs from the superposed input signal, we decode from the calibration curves that we obtained in Section 2.1.4. Completely analogously to calibration, the model simulates the MSO and lLSO membrane potentials for the input signal. The main difference is, that in the calibration procedure, the ITD was an overt input to the model and thus the assignment ITD $\mapsto$ MPA was a unique, deterministic mapping. Now the ITDs of the various individual input signals are not overtly accessible to the model and thus each ITD is covertly bound to the superposed input signal because within it the different input ITDs are contained. Therefore, we now apply the reverse operation MPA $\mapsto$ ITD where the mean population angle gives us exactly one ITD. Instead of computing an MPA (and thus an ITD) over the whole stimulus duration t, we estimate ITDs in successive time bins of length δt and the amount of ITD estimations is thus equal to the amount of time bins. Each of these time bin estimations is referred to as a *short-term ITD estimation*. Since a model assumption is that the short-term ITD estimations are considered to be performed by each hemisphere individually, a short-term ITD is retrieved from left and right hemisphere individually in each time bin. Subsequently, in every time bin these two estimates are combined. To this end we use the concept of hemispheric balancing as described in Section 1.4.1 (also cf. Lingner et al., 2018). For each time bin and each hemisphere, we denote the short-term ITD estimation as $\phi_{L/R}$ (L: left hemisphere, R: right hemisphere). These are then hemispherically balanced according to

$$\mathrm{ITD_{est}} = \arg[g(a_L)\exp(i\phi_L) + g(a_R)\exp(i\phi_R)]$$

in which $a_{L/R}$ (L: left hemisphere, R: right hemisphere) is the sum of the MSO and lLSO membrane potentials in the current time bin and $g(a) = \exp(a/a_0)$ a monotonically increasing exponential weighting function. Numerical simulations were performed for $a_0 = 0.5021$. The underlying idea is, that the higher the activity in one hemisphere, the higher the influence of the estimation of said hemisphere is on the balanced estimate and vice versa, if the activity is very low, it barely has any influence on the balanced estimate. The chosen value for a was based on the assumption that the model should always perform a clear favoring of one hemisphere even when there are only minute differences in the membrane potentials of the two hemispheres. These can then be accurately detected and the ITD estimations can be weighed accordingly. This choice for a yielded more reliable balanced ITD estimations as compared to the value $a = 0.07$ as used in Lingner et al. (2018), which is due to the increased steepness of the exponential for smaller a. The need for a different value is due to the fitted values of the BF-specific amplification factors A_{BF} in the model calibration (cf. Section 2.1.4). In case one of the hemispheres cannot perform an estimation (this is the case when there is no mapping MPA $\not\mapsto$ ITD), this hemisphere is exempt from the balanced estimation. If both hemispheres cannot retrieve an estimation, no localization of sound can be performed for this time bin. In Figure 2.7, we describe the influence of the length of the short-term ITD estimation window δt on the accuracy of the estimated short-term ITDs. To this end, we use four pure tones and two natural stimuli, all shifted by the same ITD. In each BF we vary the window δt from $1\mathrm{ms}$ to $10\mathrm{ms}$ and for each choice of δt we compute the root-mean-square error (RMSE) of all estimated balanced ITDs compared with the ground-truth ITD. We find that the value of RMSE is stable for all tested sounds for $\delta t = 10\mathrm{ms}$ (and higher) across all BF-channels. Even though the RMSE becomes worse for higher BF ($\geqslant 1124\mathrm{Hz}$), there are no more strong fluctuations for this value. Only considering pure tones, the RMSE is already stable for approximately $\delta t = 2\mathrm{ms}$ (and higher). Considering all tones, the higher the BF, the higher the localization error, which is especially noticeable for complex tones. There is therefore a dependency on low-frequency contributions for good ITD estimations. Generally, complex tones have a worse RMSE than pure tones. This is an interesting finding, because it seemingly contradicts the consistent finding in the literature, that the sensitivity to low-frequency ITDs in humans is more precise with tone complexes rather than with pure tones (Klumpp and Eady, 1956; Thavam and Dietz, 2019). McFadden and Pasanen (1976) give a possible explanation which could explain this conundrum. They point out, that different tone types have different ITD types embedded in the tone structure. Pure tones below $\leqslant 1500\text{Hz}$ have an (1) *onset* ITD and if the sound is longer than 1ms it also has an additional (2) *ongoing* ITD. For complex tones there is an additional time difference, the (3) *envelope* ITD. For low-frequency complex tones, all three types of ITDs can be used to perform lateralization, whereas for higher frequencies the ongoing ITD (2) cannot be used (cf. Figure 2.5 and the shortened areas of clear ITD resolution for pure tones at high frequencies). The missing of (2) could therefore explain why the RMSE becomes so high for BFs $\geqslant 1124\text{Hz}$. To investigate the fact, that the RMSE is higher for tone complexes than for pure tones in this model (which it should not be), all three of these ITD types should be incorporated in further studies to guarantee more consistent results with human psychophysics, as onset ITDs and envelope ITDs have been omitted from these analyses.

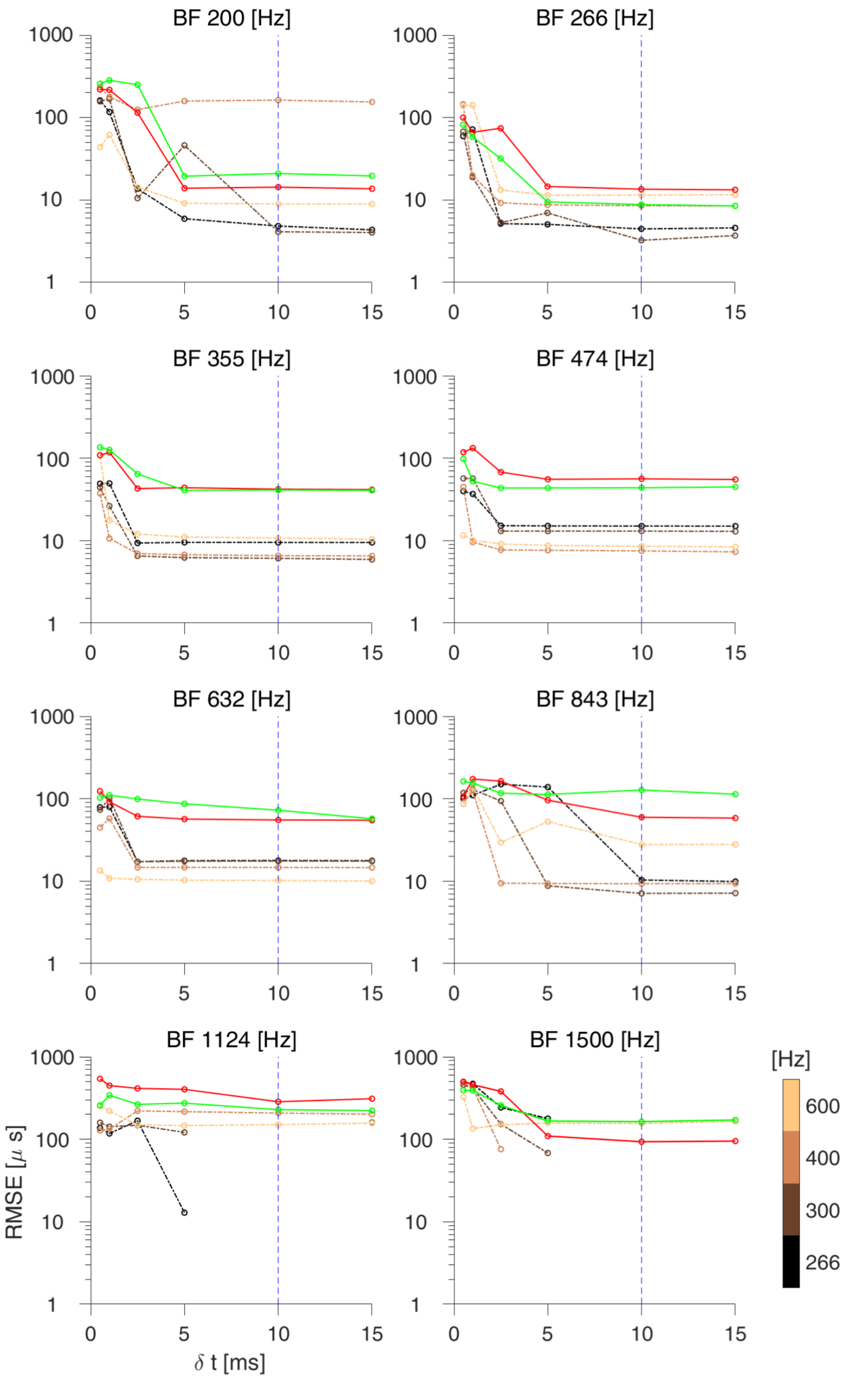

Figure 2.7: Minimizing the RMSE of short-term ITD estimation. RMSE vs. window length δt. Every panel displays the currently observed BF-channel. The colorbar shows the frequency for pure tones (dashed lines). Red and green (solid) lines show RMSE of complex stimuli (green: female speaker; red: calibration stimulus excerpt). Every signal is shifted by ITD $= -200\mu s$. For fixed window lengths δt, the RMSE is computed for all balanced ITD estimations vs. the ground-truth ITD. The vertical blue dashed line (10ms) resembles the value of δt where there are no more strong fluctuations of the RMSE for a given stimulus.

2.2.2 *Sound Source Separation*

The next subsection deals with the question, whether sound sources can be separated solely on the basis of one sound cue, the short-term ITD estimation. A successful separation of sounds only based on ITDs and neglecting other cues such as spectral components would establish a novel concept of sound source separation. To achieve this, the model must first establish how many signals are contained within the superposed signal, i.e., how many speakers are within the complex auditory scene. This number is equivalent to the number of ITDs contained within the superposed signal (see Figure 2.6). Secondly, an ITD estimation for the entire duration of the stimulus is needed, since the separation and then reconstruction of sounds will be based on these ITD estimations. The ITD estimations which are retrieved from the entire stimulus duration are now coined *long-term ITDs*. It should already be noted at this point, to prevent confusion, that long-term ITDs are directly derived from short-term ITDs and thus the latter serve as the basis of sound source separation. In other words, the long-term ITD is calculated for a time interval which is larger to the short-term ITD time interval and thus a long-term ITD stems from the distribution of short-term ITDs. To retrieve a long-term ITD estimation for a given superposed signal, we make use of the short-term ITD distributions estimated in Section 2.2.1. Short-term ITD distributions are calculated with the MATLAB built-in *histogram* function. The binning of the histogram is set to $10\mu s$. This choice is based on psychoacoustic experimentally derived value for the just noticeable difference (JND), i.e., the lowest possible ITD detection threshold, in human on average (Klumpp and Eady, 1956; Zwislocki and Feldman, 1956; Mills, 1958), even though values as low as 3.3µs JND have also been reported for individuals (Thavam and Dietz, 2019). To all histograms derived from one superposed signal, we fit a Gaussian mixture model (GMM) distribution which is used to estimate the long-term ITDs.

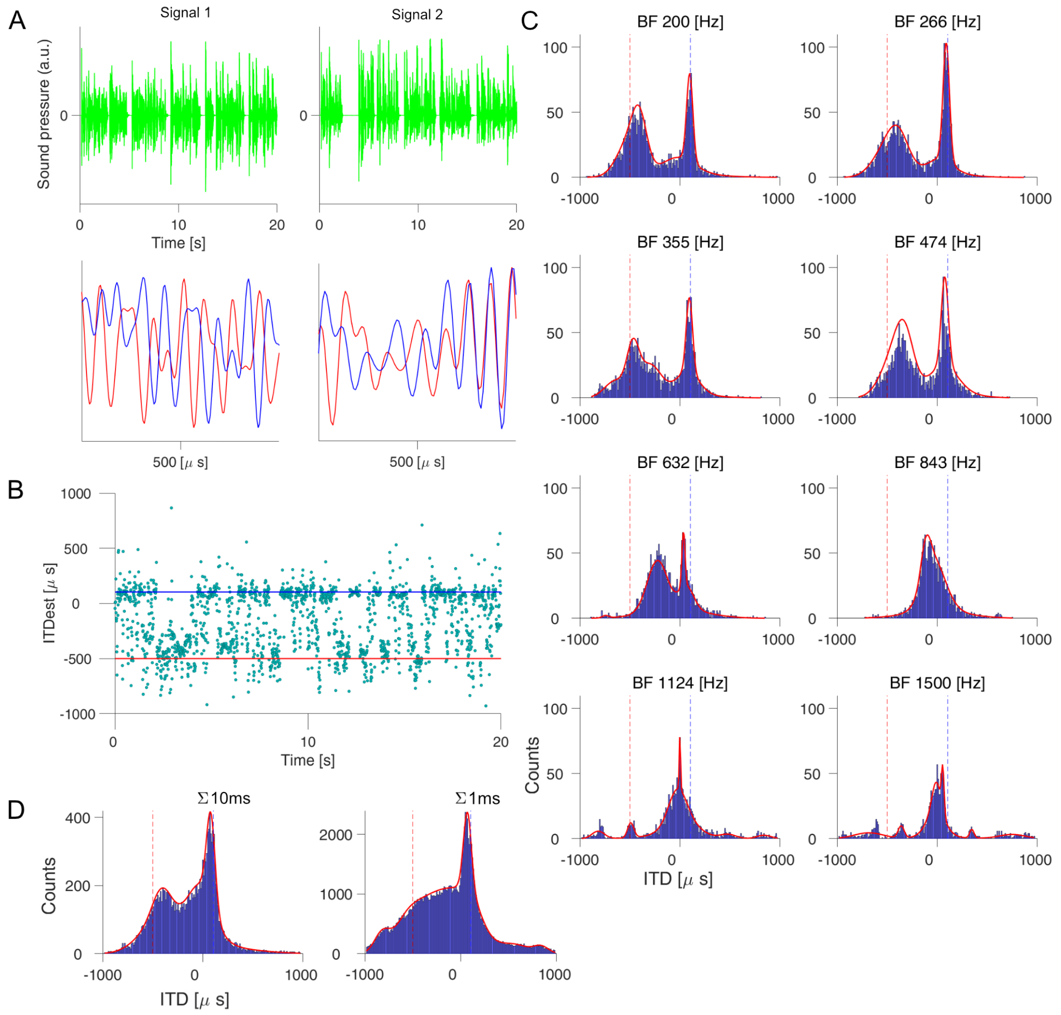

Figure 2.8: Short-term ITD estimation and GMMs. A. Two input signals (green graphs), male and female, shifted by $\text{ITD}_1 = -500\mu s$ and $\text{ITD}_2 = +110\mu s$, respectively). Bottom Graphs are a close-up where red shows the arrival at right ear and blue at left ear. B. Short-term ITD estimation every $10ms$ for the sum of the inputs in A in the 200Hz BF-channel. Red and blue horizontal lines are ground-truth ITDs. C. Fitting of GMMs for $10ms$ short-term ITD distributions in each BF-channel for the entire stimulus duration. Histogram binning is $10\mu s$. Red and blue vertical lines are ground-truth ITDs. D. Sum histograms of all BF-channels in C. Left histogram is for $10ms$, right for $1ms$ short-term ITDs. See main text for further description.

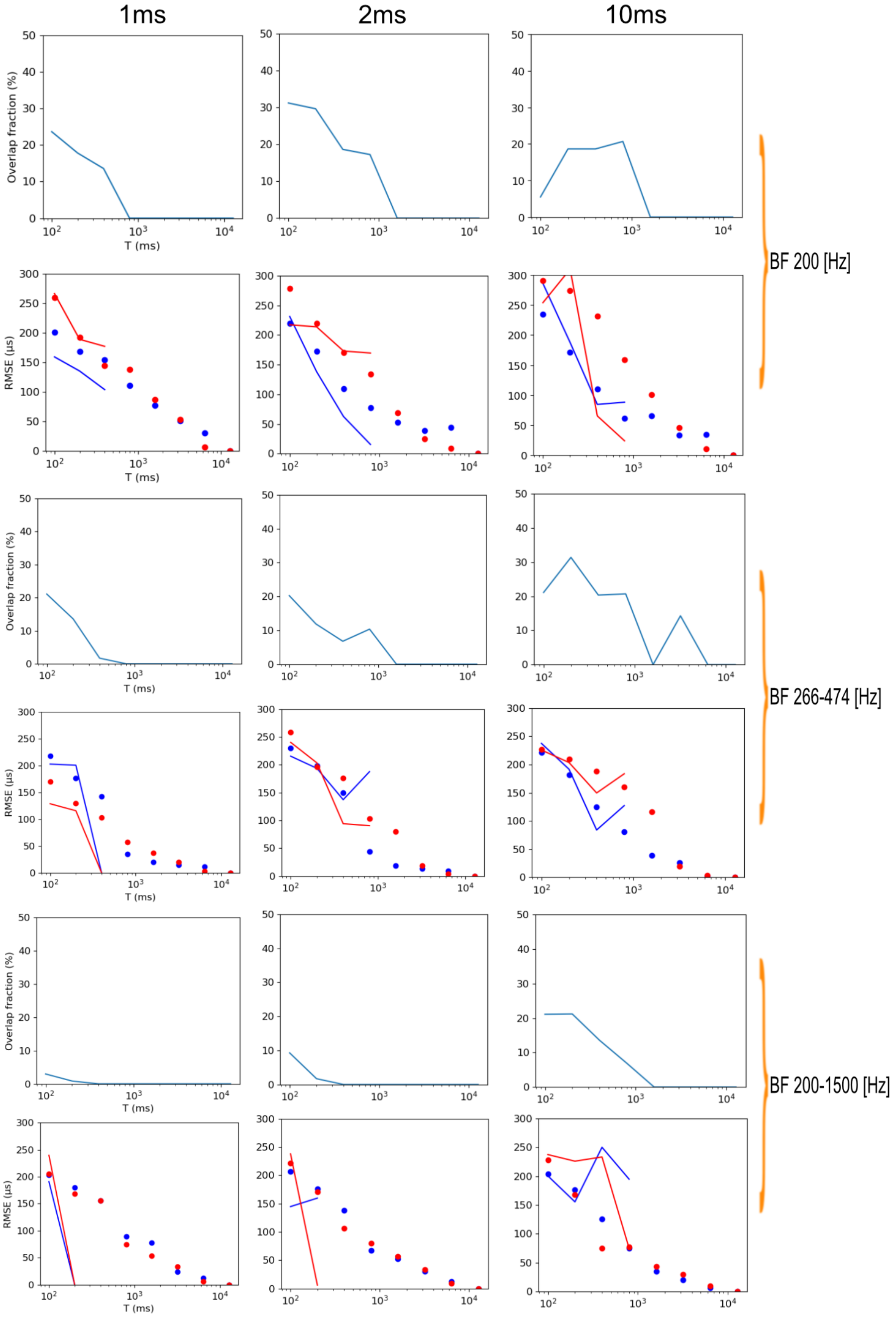

Figure 2.9: The RMSE vanishes for sufficiently long-term ITD estimations. For a fixed short-term ITD window length (columns; 1ms, 2ms and 10ms) and for different frequency bands (indicated next to orange arrows) the long-term ITD window length T is fixed. For every T, a GMM is fit to the distribution of short-term ITDs to retrieve an ITD estimate for each speaker. Uneven rows: Overlap fraction vs. T. Even rows: RMSE vs. long-term window length T. Ground truth ITDs: $\mathrm{ITD}_1 = -500\mu s$ (red), $\mathrm{ITD}_2 = +110\mu s$ (blue). Solid lines indicate when the GMM fit the amount of speakers via the BIC. Dots indicate when the number of GMM components was set to two. See main text for further explanation.

Gaussian Mixture Models and Long-Term ITDs. A GMM makes the assumption, that all data points are produced from of a mixture of k ($k \in \mathbb{N}_{\geqslant 1}$) Gaussian distributions. The parameters of these Gaussians are unknown. The use of GMMs is a straightforward choice for our underlying data, because we observed most of the short-term ITDs to scatter mainly around the ground-truth ITD and few wrong estimates to deviate further from the ground-truth ITD which precisely fits the distribution of a Gaussian. Considering our auditory scene analysis, this means that if it simply consists of only one signal, then we would also expect only one Gaussian to be fitted to the data and its peak to correspond to the ground-truth input ITD. If the superposed signal is complex and made up of $2, 3, ..., n$ input signals, we would also expect $2, 3, ..., n$ Gaussians, respectively, with each peak approximating one ground-truth ITD of the individual input signals. Therefore, the number of Gaussian curves or components of the GMM can be thought of as representing the number of speakers of the complex auditory scene. However, there are cases where there are more components fitted than speakers are contained within the signal, but this is only the case if the peak of one Gaussian G_1 is under the curve of another fitted Gaussian G_2 and if the means are close to each other. We use the object *GaussianMixture* implemented in Python Scikit-learn, which employs an iterative Expectation-Maximization (EM) algorithm to maximize the likelihood of each Gaussian. From the superposed input signal, this procedure extracts the amount of speakers (i.e., the number of maxima) as well as each long-term ITD which is equivalent to the means of every fitted Gaussian. A problem that can arise is that the number of Gaussians that the GMM is trying to fit to the data is too large, i.e., there are too many components observed in the GMM. The correct number of components, i.e., the correct amount of Gaussians fitted to the data is of high importance, because it directly reflects the numbers of speakers within the complex auditory scene. To obtain a reliable estimate for the number of components, we calculate the Bayesian information criterion (BIC) which, in short, seeks the best-fit model to accurately represent the underlying data (Schwarz, 1978; Wit et al., 2012). The best-fit model is the one that has the least information loss when compared to other models. The BIC is calculated for each GMM with number of components ranging from 1 to 5 in increments of 1. Choosing the model with the lowest BIC results in the best candidate for the number of components. The number of components retrieved from the BIC were compared to the similar Akaike information criterion (AIC) (Akaike, 1973; Akaike 1974). Calculation of the AIC delivered no differences to the number of components when compared to BIC.

Using the GMMs, we can now analyze the influence of the long-term ITD window length T which is necessary for a good fit of Gaussians in order to retrieve ITDs that do not deviate too much from the ground-truth ITDs. The exact procedure and the derived GMMs from all BF-channels are shown in Figure 2.8. Two input signals (green graphs; male and female speaker, 20s stimulus) are shifted by $\mathrm{ITD}_1 = -500\mu s$ and $\mathrm{ITD}_2 = +110\mu s$. The bottom graphs are a close-up (2.8A). Red resembles the input arriving at the right ear, blue left ear. For the sum of the time-shifted input signals of A, balanced short-term ITD estimations (green circles) are performed for the whole stimulus duration in time bins of $10ms$ (2.8 B). Blue and red horizontal lines are ground-truth ITDs. The estimations temporally cluster around different ITDs for different time windows in line with the theory of glimpsing (cf. Cooke, 2003), where a signal of a complex scene can be perceived more dominantly than another for a brief period of time before another signal is perceived more dominantly. For the whole stimulus duration and to each BF-channel, a GMM (red curve) is fitted to the balanced short-term ITD estimations grouped into histograms of $10\mu s$ (2.8C). The location of the stimuli is estimated by the maxima of the Gaussians. For higher frequency ($\geqslant 843Hz$) channels, the estimations become less reliable and deviate more from the ground-truth ITDs. The two GMMs in 2.8D show the short-term ITD estimations summed up for all frequency channels. Left for $10ms$ short-term ITDs and right for $1ms$. It is noticeable, that the sum histogram for $1ms$ has worse ITD estimations than the sum histogram for $10ms$. We consider this to be the result of the higher BF-channels, which generally have worse estimations than the low-frequency channels. To this end, we therefore compared the RMSE of long-term ITDs calculated for varying window lengths T in different frequency bands (cf. Figure 2.9). For fixed short-term ITD window lengths of $1ms$, $2ms$ and $10ms$, we calculate the RMSE under two different conditions. In the first condition, the GMM was supposed to gauge the correct number of speakers, in the second condition, the GMM was set to estimate ITDs with two fixed components, i.e., the GMM knew *a priori* that there were two speakers. The amount of correct detections of the number of stimuli over time bins is defined as the overlap

fraction. An overlap fraction converging to zero implies that the GMM does not detect the correct amount of signals anymore. For sufficient long time bins T, this case always occurs. This is in line with the result that the sum histogram for $1ms$ has worse estimations than for $10ms$ as the overlap fraction is much lower and converges faster to zero than for $10ms$.
For the other condition that the number of components was set to two, it is apparent that the longer the long-term ITD estimation is, the more precise the estimation gets. In every case, for sufficient long times T, the RMSE converges to zero which means a perfect estimation. From both conditions we conclude, that there are two questions to be answered when performing sound localization. The first question is how many stimuli are present in the auditory scene. The second question is where are the stimuli. For long time windows T, the precise gauging of numbers of speakers becomes more difficult, i.e., for very short time windows T an estimate of the number of speakers can be extracted. While lateralization works well in any case (i.e., a general idea where the stimuli are located), the precise estimation of stimuli takes at least $1-2$ seconds, depending on the stimulus. It is therefore not unreasonable to assume, that these two processes are carried out subsequently when trying to localize sounds. First, the number of speakers is estimated and when this number is known (corresponding to setting the GMM to two components), then the reliable estimation of the direction of the sound sources can be carried out.

The Moore-Penrose Pseudoinverse. Once the ground-truth ITDs of the complex auditory scenes have been estimated via the GMMs, the next step is to successfully separate and reconstruct the original sounds. To demonstrate the underlying idea, we first restrict our complex auditory scene to two signals S_1 and S_2 with corresponding ITDs ϕ_1 and ϕ_2, respectively. The superposed signals $S_{R/L}$ arriving at the right and left ear are then given by the following two equations.

$$S_R(t) = S_1(t + \tfrac{\phi_1}{2}) + S_2(t + \tfrac{\phi_2}{2})$$

$$S_L(t) = S_1(t - \tfrac{\phi_1}{2}) + S_2(t - \tfrac{\phi_2}{2})$$

We desire an expression only dependent on one signal. Thus, we can make the subtraction ansatz

$$\Phi_1(t) = S_R(t-\frac{\phi_2}{2}) - S_L(t-\frac{\phi_2}{2}) = S_1(t+\frac{\phi_1}{2}-\frac{\phi_2}{2}) - S_1(t+\frac{\phi_1}{2}-\frac{\phi_2}{2})$$

which is independent of S_2. Computing the Fourier transform gives us

$$\widehat{\Phi}_1(\omega) = \widehat{S}_1(\omega) \cdot 2i\sin\left(\frac{\phi_1-\phi_2}{2}\omega\right)$$

which can now be easily solved for $\widehat{S}_1(\omega)$. Computing the back transform results an explicit solution for $S_1(t)$. Analogously, this method yields an explicit solution for $S_2(t)$.

Obtaining an expression only depending on one signal is unfortunately not so simple when regarding complex auditory scenes with more than two input signals. The superposed signals $S_{R/L}$ arriving at the right and left ear for n signals are then given by:

$$S_R(t) = S_1(t + \frac{\phi_1}{2}) + S_2(t + \frac{\phi_2}{2}) + ... + S_n(t + \frac{\phi_n}{2})$$

$$S_L(t) = S_1(t - \frac{\phi_1}{2}) + S_2(t - \frac{\phi_2}{2}) + ... + S_n(t - \frac{\phi_n}{2})$$

From these equations, it is impossible to find a subtraction ansatz so that the resulting expression is only dependent on one speaker, i.e., the resulting equations will always be dependent on $n-1$ input signals. In terms of linear algebra, for n signals, we receive a $n-1$-dimensional system of linear equations with n variables. Such a system of linear equations does not have a unique solution as its coefficient matrix is singular and thus has no inverse.

A remedy is that every system of linear equations $A\vec{x} = \vec{b}$ can be solved with the Moore-Penrose pseudoinverse A^+ which is always existent and unique (Moore, 1920; Penrose, 1955). In short, the pseudoinverse minimizes the problem $||A\vec{x}' - \vec{b}||$ where $||\cdot||$ describes the Eucledian norm. The vector $\vec{x}'$ then represents the best approximation to the solution of $A\vec{x} = \vec{b}$. The pseudoinverse can be shown to be defined as

$$A^+ = \lim_{\epsilon\to 0}\left((A^TA + E\epsilon)^{-1}A^T\right)$$

where A^T denotes the transpose of A and E is the identity matrix of dimension $\dim(A^HA)$. For our calculations, we set the regularization factor $\epsilon = 0.001$. This value for ϵ resulted in sound reconstructions with a correlation of 1 (original vs. reconstructed signals) when the estimated long-term ITD was identical to the ground-truth ITD (cf. next subsection). The pseudoinverse thus now provides a means of sound source separation and sound reconstruction when there are two or more input signals in the complex auditory scene.

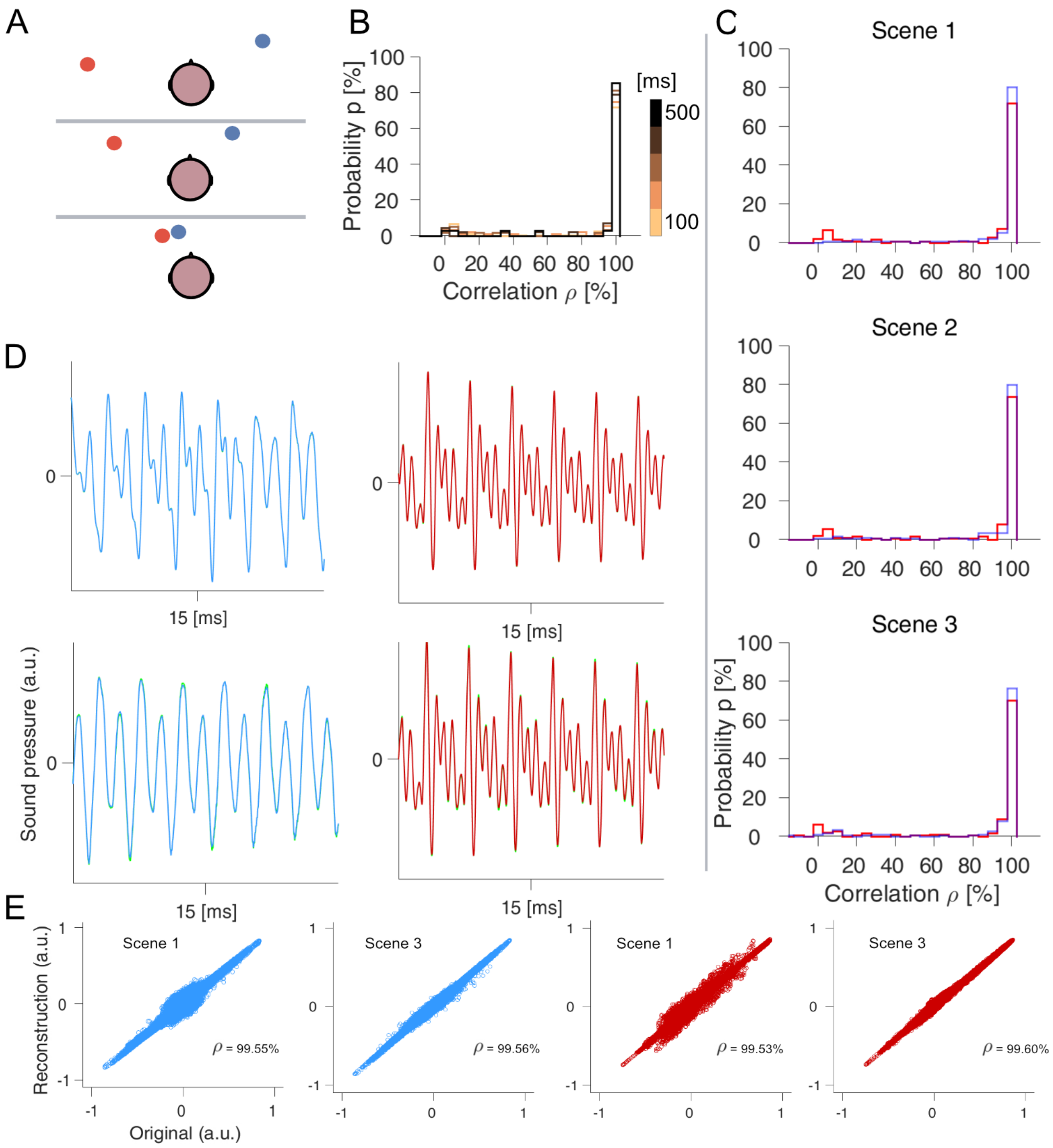

Figure 2.10: Sound separation and reconstruction. A. Three auditory scenes with two signals at different azimuthal positions. Scene 1 (top): $\mathrm{ITD}_1 = -500\mu s$, $\mathrm{ITD}_2 = +110\mu s$. Scene 2 (middle): $\mathrm{ITD}_1 = -404\mu s$, $\mathrm{ITD}_2 = +81\mu s$. Scene 3 (bottom): $\mathrm{ITD}_1 = -220\mu s$, $\mathrm{ITD}_2 = -95\mu s$. B. Probability vs. correlation. Correlation and occurence was measured for different reconstruction intervals (color code). C. Probability vs. correlation in the three different scenes from A for a reconstruction interval of 500ms. D. Original signals (green) shown with their reconstruction (blue, red). Top panels scene 1, bottom scene 3. E. Reconstruction vs. original signal scatter plots of the corresponding signals in D. Values close to identity line represent good reconstructions. The correlation coefficient for the entire stimulus duration is indicated by ρ.

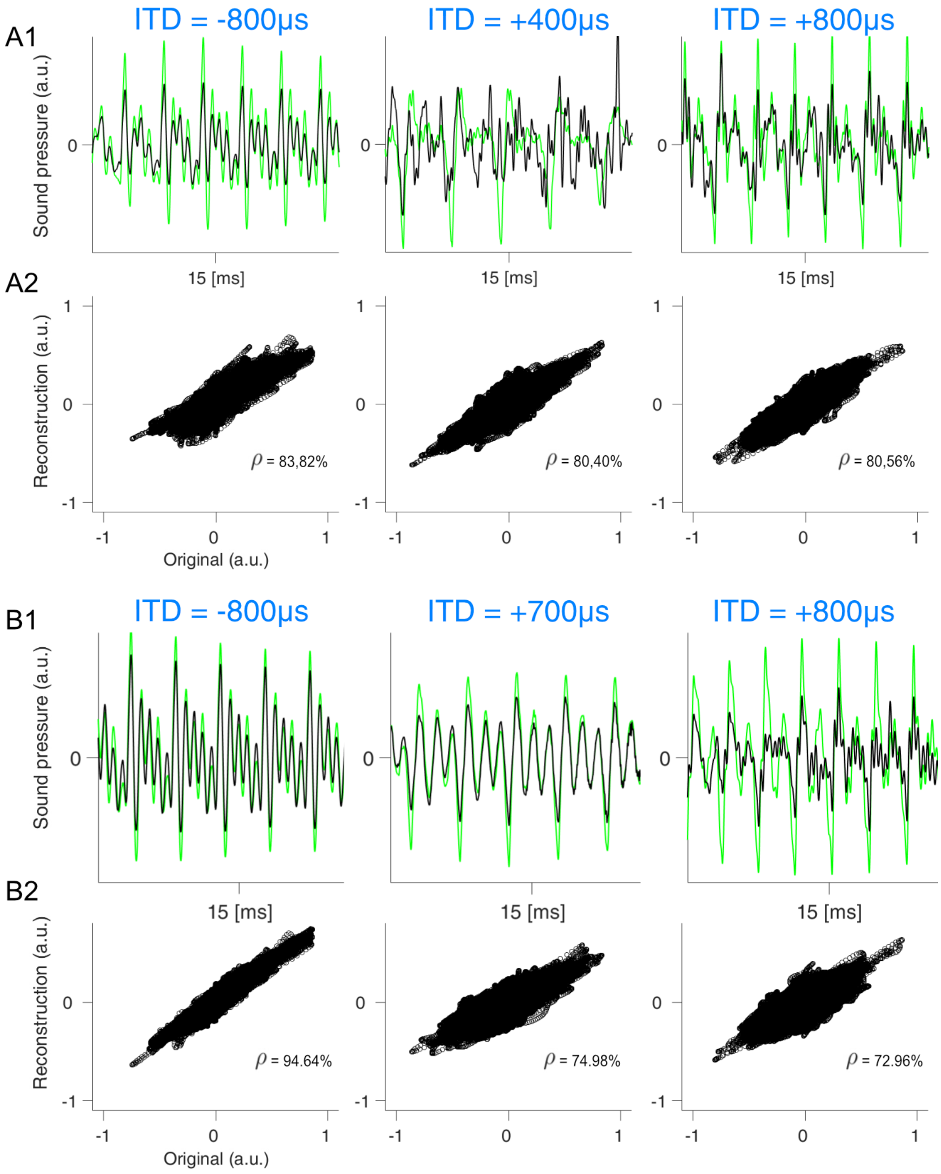

Figure 2.11: Separation of three signals in two different auditory scenes. Group A. The three speakers are positioned at $\text{ITD}_1 = -800\mu s$, $\text{ITD}_2 = 400\mu s$ and $\text{ITD}_3 = 800\mu s$. ITD indicated in blue. Group B. The three speakers are positioned at $\text{ITD}_1 = -800\mu s$, $\text{ITD}_2 = 700\mu s$ and $\text{ITD}_3 = 800\mu s$. ITD indicated in blue. A1/B1. Three different signals (left to right panels). Original signals in green, reconstructions in black. A2/B2. Reconstruction vs. original signal scatter plots of the corresponding signals (from left to right) in A1/B1. Values close to identity line represent good reconstructions. The correlation coefficient for the entire stimulus duration is indicated by ρ. See main text for further description.

Separation and Reconstruction. In the final steps, we now apply the Moore-Penrose pseudoinverse to a variety of complex auditory scenes to obtain separated estimated input sounds from the original superposed input signal. To recapitulate, the superposed input signals $S_{R/L}$ at right and left ear consist of n signals S_n shifted accordingly to their individual ground-truth interaural time differences $\frac{ITD_n}{2}$. The GMMs estimate the amount of input signals and an estimator of the azimuthal position of each input signal in the form of long-term ITDs on the basis of short-term ITDs. These estimated ITDs are denoted by ϕ_n. The Moore-Penrose pseudoinverse uses these estimators ϕ_n to separate the superposed input signal. Each separated sound retrieved via the pseudoinverse is thus itself an estimator $\widehat{S}_n$ to each corresponding input signal S_n. Therefore each $\widehat{S}_n$ is a reconstruction of the original signals.

To measure quality of reconstruction, we have developed three auditory scenes (2.10A) based on the analysis from the long-term ITD estimations (cf. 2.8). All scenes have the underlying ground-truth ITDs of speaker one with $ITD_1 = -500\mu s$ and speaker two with $ITD_2 = +110\mu s$ (male and female speaker, respectively, each $20s$ stimulus). In the first scene (2.10A, top), we assume that the estimated ITDs (red and blue dot) are equivalent to the ground-truth ITDs. In the second scene (2.10A, middle), we assume that the estimated ITDs are perceived distorted, i.e., pushed towards the midline by the $5\%-$quantile RMSE of the short-term ITDs over the whole stimulus ($96\mu s$ for ITD_1 and $29\mu s$ for ITD_2). In the third scene (2.10A, bottom), we assume the estimated ITDs to be pushed towards midline by the $95\%-$quantile of all RMSE calculations ($280\mu s$ for ITD_1 and $205\mu s$ for ITD_2). To find a suitable binning window for reconstruction, we tested scene 1 for varying reconstruction window lengths for signal 1 (Figure 2.10B), color code resembles different window lengths. We then use the MATLAB function *corrcoef* to measure the correlation ρ between original and reconstructed signal in every time bin. A perfect correlation is denoted by $\rho = 100\%$ and no correlation by $\rho = 0\%$. The latter is defined as the null hypothesis H_0 that there is no statistically significant relationship between original and reconstruction. Each ρ determined by *corrcoef* has a corresponding p-value which is the probability of H_0 being correct. The significance level is set to $\alpha = 0.05$. A p-value of exactly 0.05 would entail that the determined ρ is due to only 5% chance. We reject the null hypothesis if the p-value is smaller than α and then favor the alternative hypothesis, i.e., that there is a statistically significant correlation between original and reconstruction (H_1). Independent of reconstruction interval length, we found the most reliable reconstruction for pseudoinverses for the regularization factor $\epsilon = 0.001$. Furthermore, we found that larger window lengths show a higher probability of perfect reconstruction (Figure 2.10B, correlation histogram binning in 5% increments from), therefore we set the reconstruction interval to $500ms$ for the following analyses. Figure 2.10C shows the correlation of the separated and reconstructed signals (red ITD_1, blue ITD_2). The closer the signals move to each other, the more difficult separation becomes (scene 3). In the initial condition (scene 1) separation for both signals works very well for having a perfect reconstruction (over 95%) for over 75% of the signal. A close up (2.10D, top two panels) of the reconstruction signals (blue and red graphs) in scene 1 shows how good the resemblance of the input stimulus (green graph) is. For scene 3 (bottom two panels), the resemblance is lost. This is probably due to the fact, that in this scene the two input signals are of almost overlapping origin. In Figure 2.10E, we show the original vs. the reconstructed signal. For a perfect reconstruction, the data points would all be on the identity line. Thus, less deviations from the identity line resemble better reconstructions.

A further example of sound separation and reconstruction of three stimuli is shown in Figure 2.11 for two different auditory scenes. In scene 1 (group A), the three speakers are positioned at $ITD_1 = -800\mu s$, $ITD_2 = 400\mu s$ and $ITD_3 = 800\mu s$. In scene 2 (group B), the three speakers are positioned at $ITD_1 = -800\mu s$, $ITD_2 = 700\mu s$ and $ITD_3 = 800\mu s$. Thus, in the second scene, the second speaker has been moved away from the third speaker closer to midline. In general the reconstructions are not as good as compared to the scenario when only two speakers are present (Figure 2.10), but it is apparent, that the more distance the speakers have to each other, the better their own reconstruction becomes.

Taking all results into consideration, we conclude that it is theoretically possible, that the short-term ITD can serve as the sole cue to localize individual sounds in the azimuthal plane in complex auditory scenes. Furthermore, it can also theoretically serve as the sole cue to separate and reconstruct sounds from the same complex auditory scene, given that the estimated ITDs do not deviate too far from the ground-truth ITDs.

3 DISCUSSION

... but some are useful.

— George E. P. Box

3.1 SUMMARY OF RESULTS

In this thesis we have sought to satisfy three main objectives in order to obtain insight about a population code for sound localization in horizontal space. In a first objective, we provided the basis of this work, namely a phenomenological effective model of binaural hearing in humans. With the help of this model, we provided a possible neuronal code for the spatial representation of acoustic space. In a second objective, the feasibility of this code was tested by applying it to real-life speech to test whether it can be utilized to estimate ITDs. In a third objective, we tested if the code could be used to separate sounds from each other in complex auditory scenes and, finally, if they could be reconstructed. We now recapitulate the main results from these three aims.

1. The Effective Model.
   The neuronal population code itself is the result of our approach towards an effective model, which was designed such that it is fully biologically motivated and should faithfully mirror large parts of the phenomenology of the human binaural hearing apparatus. To this end, we have implemented the two ITD-sensitive nuclei MSO and lLSO. We have bypassed an implementation of the complex SOC circuits by assigning to every neuron three defining characteristics: the BF, CD and CP. An arbitrary binaural pressure wave (a signal) was then assigned to a fixed ITD and run through standard peripheral and binaural processing. The resulting output was the membrane potential for each

of the two nuclei. From the two-dimensional space that these two membrane potentials span, we have derived a neuronal code of ITDs. Since the two nuclei were each considered in their entirety, i.e., all cells participated in the establishment of the code, the ITD encoding mechanism is thus based on a population code. The neuronal realization (i.e., the population response) of an ITD is the MPA. To every BF and to every ITD such an MPA exists and consequently an ITD can be encoded by its corresponding MPA. To guarantee a stable ITD encoding mechanism, we have calibrated the model with a sufficient long stimulus which spans a wide range of frequencies. Finally, the reverse operation of inferring an ITD from the MPA to a fixed BF for a given arbitrary complex signal was tested and yielded robust results (correct estimations). The neuronal population code we proposed for sound localization in the azimuthal plane is thus fully captured by one single quantity, the MPA.

2. Short-Term ITD Estimation.
   The effective model itself and the newly established population code for auditory space were then applied to estimate ITDs in cocktail party scenarios. This was simulated by summing up n input signals at different azimuth. Application of the effective model to the superposed input signals yielded the base raw data for the ITD estimation procedure, namely the membrane potential of the complex signal in 2d-space (lLSO $V(t)$ vs. MSO $V(t)$). From the MPAs we then decoded the corresponding ITDs of the input signals. For biological plausibility, the individual estimations of ITDs were performed in short time bins δt. Each estimation in such a time bin was coined a short-term ITD. The short-term ITDs in each hemisphere were then balanced according to their responses in the two ITD-sensitive nuclei. Applying hemispherically balanced short-term ITDs to complex auditory scenes gave rise to the observation of the phenomenon of glimpsing, i.e., the temporally restricted clustering of multiple subsequent short-term ITD estimations close to one signal before switching to another signal. This means that a listener would perceive the utterance of one speaker more dominantly before perceiving the utterance of another speaker more clearly. Lastly, in accordance with the phase

locking behavior of auditory nerve fibers accompanied by the high fidelity relay of APs from bushy cells to MSO and lLSO neurons and the duration restrictions it presupposes on the time window of synaptic integration of the two ITD-sensitive nuclei, we observed a BF-dependent worsening of the ITD estimations in case the length of the time bin δt is only $1-2ms$ which is still low enough to allow for successful integration of synaptic inputs (see 3.3.1) when considering that long-term ITD estimations (which themselves contain a specific amount of short-term ITDs and are only a few hundreds of microseconds long; see below) can be seen as an analogue to the integration process.

3. Sound Separation.
In the final objective, we analyzed if the neuronal representation of auditory space, as supplied by the model in the form of the MPA yielding a specific and unique ITD, could serve as the sole basis for sound source separation. To this end, we introduced the concept of long-term ITDs which were themselves based on short-term ITDs. The latter were extracted from the information contained within the population response (i.e., the value of the MPA) and subsequently grouped together by corresponding histograms. GMMs were then used to extract the number of original input signals contained within the superposed signal, i.e., the complex auditory scene. The means calculated by the GMMs represented the aforementioned long-term ITDs which were interpreted as the ground-truth ITDs of the various input signals. Thus, at this stage, the azimuthal allocation of all speakers was completed. The long-term ITDs then served as the basis of the sound separation procedure. In order to isolate each input signal from the complex signal, the Moore-Penrose pseudoinverse could be used as it provided a means of solving the underlying system of linear equations which arose when summing up two or more time-shifted signals. The separated sounds were then reconstructed. This was not performed for the entire complex input stimulus duration at once, but for smaller subsequent reconstruction intervals. We found that for reconstruction intervals of $500ms$ the correlation with the original signals were over 95% for over 75% of the entire stimulus duration ($20s$), given that the estimated ITDs are close enough to the ground truth IDs. We finally

concluded our proof of principle, namely that the ITD could theoretically serve as the only cue to separate and reconstruct up to three sound sources in horizontal space when they are contained within an auditory scene.

3.2 MODELING: CHANCES AND LIMITATIONS

3.2.1 *Models and Reality*

All of the results presented in this thesis are the direct product of a theoretical computational model. The modeler must decide on which biological aspects to focus on more in-depth and which aspects to cover more superficially. This, of course, can only happen if the modeler moves within a plausible and realistic framework that does not give a distorted depiction of the real world we live in. The fact that it is simply impossible to integrate every detail of reality into a model has the immediate implication that all models of biological systems will inevitably be missing some features of the overall picture. This means that the features of a model and all the results it entails can never be a direct 1-to-1 mapping of the real world. In other words, a model will always be wrong or faulty in the sense that it will never be 100% complete. Nevertheless, models are useful tools to help us understand at least pieces of how a biological system might work and therefore we delineate the general chances that go hand in hand with the methodology of modeling, but also scrutinize the restrictions and dangers the modeler is confronted with when designing such a model and put them into perspective with the current model at hand.

Advantages of Modeling. One of the biggest advantages of modeling is that computational models can always be produced to describe biological systems without the actual biological system having to be readily available (in form of test subjects, brain slices, etc.). Models are thus easy to handle tools to facilitate the understanding of phenomena and they can provide novel insights into neurobiological aspects that would otherwise be difficult to test for. The reason for this is that the complexity of models and the model itself is always under full control of the modeler. This means that the amount of variables is precisely set and they can be either reduced or increased to investigate different scenarios. Needless to say, the variables in real life systems cannot be controlled with such precision. Having full control over a model results in reproducibility of results which makes it possible to analyze causalities in different biological settings by changing parameters and variables. For example, in our model it is very easy to shift the location of the input sounds in space since this is just a number (the ITD) that we can manipulate. There is no upper bound (except CPU hours) that restricts us on the amount of different auditory scenarios that we can create. Compared to an on-site experiment, there are, however, many limiting factors, such as time, precise speaker placement and limited attention spans of subjects.

A good model should furthermore be simple. Not only because this simplicity allows for the identification of said causal components, but also because the simplification process itself already identifies which key aspects must be included in the model and which aspects can be excluded. Our model is simple in the sense, for example, that in the neuron population implementation we only consider their defining properties (BF, CD and CP) and do not consider the entire SOC circuits.

Finally, models are also highly informative when they produce negative results (i.e., the hypothesis being tested is not verified) or if the model breaks down altogether. Because this does not mean that the model is wrong *per se*, it can simply mean that the model must be altered in its current form and its configurations to accurately capture the biological system or experiment in question. Thus, the so-called breakability of a model can lead to a more precise and better developed model which, in turn, can reveal those key features of reality which are inevitable to the model and cannot be left out of the implementation. Of course these notions can also all be applied to live experiments, but again, it is the full control over the experiment settings and the simplicity that make models so advantageous as a repetition of an experiment is much more time consuming, costly and uncertain (i.e., the experiment may fail again) than tweaking features of a model. (O'Reilly and Munakata, 2000; Bray, 2014; Teufel and Fletcher, 2016). In our model, we established that using any arbitrary ITD to shift the input signals would result in unreliable calibration curves which would cause the model to break down because no reliable estimation of ITDs could be made. This could easily be fixed by only considering ITDs that were multiples of the inverse sampling rate. Although this poses a restriction on the modeling process and the ITDs that can be analyzed, it does not mean, that the real biological system cannot correctly localize other ITDs.

Problems of Modeling. The most straightforward problem of a model is that it can be too simple in comparison to the real-life scenario. It may not capture core elements of the underlying biology or the overall experiment and these elements and their implications for the results may then simply go unnoticed by the modeler. The pitfall of over-simplification may thus result in a model that is not valid or cannot explain certain underlying phenomena. A theoretical example would be if we considered the anatomy of the SOC circuits to be unknown nowadays, but the properties of the neuron population to be known. From the results of our model, we could not make any statements about the existence or importance of any underlying biological circuits. Because of the simple structure of the model, the implication for an intricate biological pathway might go unnoticed. Of course we can only speculate if all the parameters in our model are sufficient to explain the real biological system. We simply cannot know if something is missing. For example, we are able to infer ITDs from the membrane potentials of lLSO and MSO and use them for sound localization. But there could theoretically be other factors that we are missing, e.g., high plasticity or significantly varying membrane time constants of different cells.
On the other hand, however, a model may become far too complex by considering too many variables. In this case it may be that the model produces good results, but it need not be the case that the real-world biological system actually makes use of all of these variables. The biological system might perform just as well with some of the variables being absent, rendering them redundant for the entire model. Considering too many variables and thereby increasing a model's complexity is furthermore not beneficial to the modeler since it results in unnecessary high computation time. A model can quintessentially do anything the modeler desires. In theory, it is possible to implement so many degrees of freedom, that the model will always verify a given hypothesis. It is therefore possible that different models can provide different explanations for the same biological phenomenon (O'Reilly and Munakata, 2000; Teufel and Fletcher, 2016; Bray, 2014). In our case, we tried to avoid this problem by using a minimal variables approach (only considering those variables

that - if absent - would cause the model to break down), and also we have developed a model which is not restricted to the uses it is applied to in this thesis. For example, it could be applied to make predictions for psychophysical experiments where the following questions could be answered: How many speakers can be maximally resolved and at which localization success rate? When does a test subject realize how many speakers are speaking simultaneously and when does the test subject know where the amount of gauged speakers are located in azimuthal space?

Verdict. As has been demonstrated, modeling has many benefits and also some disadvantages. From these we cannot and should not decide if modeling or another methodology is more superior than the other. Rather, different methodologies can be seen as complementary to each other, i.e., models can be used to design experiments and, vice versa, experiments or the results hereof can be used as the basis for modeling. However, we can conclude that one has to be wary of the fact that each method brings along its own chances and limitations that one must be aware of to obtain results that assure scientific integrity. Considering all of the aforementioned pros and cons, we can conclude that modeling requires the precise taring of situation-tailored simplification and likewise complexification to retrieve an (at least approximately) accurate implementation of reality. To make a model scientifically relevant, one should therefore always ask what is exactly being modeled, how close does the model approximate reality and what aspects from reality are missing. Most importantly, these questions must constantly be reevaluated and put into perspective with the current state of the model (Teufel and Fletcher, 2016; Bray, 2014). In the next subsection, we turn to the most important parameters of our model and discuss them while keeping these mentioned chances and limitations of modeling in mind.

3.2.2 *Model Size*

One of the model parameters that has greatest influence on computational time, and thus being a strong limiting factor, is the size of the population. As one main objective was to implement a fast model, we set the total amount of neurons of all four nuclei to a constant number which we denote as M. The model used in this thesis is based on a population of exactly 800 MSO and 408 lLSO neurons per hemisphere, therefore $M = 2416$ which is roughly 20.58% of the total amount of neurons ($H = 11742$) estimated by Hilbig et al. (2009). The question now is, if changes in model size also significantly change the calibration curves which are used to estimate the ITDs or if they are invariant under size manipulation. In order to test for this, we calibrated the model for different sizes of M, namely $M' = M \times d$ for $d \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, ..., 5, 6\}$. Note that the factor 5 already exceeds the amount H of neurons gauged by Hilbig and 6 results in model nuclei that are considerably larger. The resulting calibration curves were compared to those that were computed with the exact number of neurons as stated in Hilbig et al. (2009). We found, that for the factor $d = \frac{1}{2}$ the maximum difference between all MPAs to each ITD (calculating the absolute value after subtracting the original calibration curve from the manipulated curve) was $\approx 1.6\,\text{deg}$. For the factors 1 and 2 it was as small as $\approx 1\,\text{deg}$. For lower factors $d \leqslant \frac{1}{2}$ the difference was larger $> 2\,\text{deg}$ and for higher factors $d \geqslant 3$ the difference also yielded ≈ 1 deg. Since a factor of $d = 2$ already resulted in computation times that were four times longer than for $d = 1$, we opted for the latter as the best trade-off between time and accuracy. Furthermore, we argue that a difference of ≈ 1 deg is completely acceptable, because this is the minimum audible angle in humans as proposed by Mills (1958). The overall similarity between the curves can be explained by the neuron population parameters CD, CP and BF, because the CD is drawn from a normal distribution (via the BP and the BF; see equation (2.3), Chapter 2) and the CP is drawn from a uniform distribution within specific limits determined by Lingner et al. (2012). The higher the number of neurons gets, the denser the population cloud gets (cf. Figure 2.1) and for $d > 3$ it does not make any significant differences in the calibration curves anymore.

We can conclude, that in case an ideal population size exists, then it will be close to, or around, H, but we have seen that a population size of M is already sufficient to obtain an ITD encoding mechanism. Which leads us to the question: Does an ideal population size exist? We argue that this may not be the case as we suggest that there might not even exist a ground truth calibration curve. It is a fact that the MPAs are more similar to the original calibration the more the size of M' approximates H. But this misses the point that the calibration curves are simply used as a means to perform ITD estimations. There is no reason to assume that the calibration curves retrieved from H represent reality better than the ones retrieved from M′. For M and H, we determined in both cases that the resulting curves could perform ITD estimations with the same precision. It is therefore not unreasonable to assume that, if the human auditory system makes use of such a calibration curve concept, then the real curves will most likely be different for each human and their precise shape is dependent on the individual nuclei sizes and probably display some sort of plasticity. This is motivated by the fact, that the calibration curves are very different for pure and complex stimuli. It nevertheless still remains a proof of principle that ITDs can be encoded via such curves, but we conclude that the ITD encoding mechanism might be invariant or at least very robust to population size.

3.2.3 *Population Membrane Potentials*

In this thesis, ITDs are estimated through the metric of population membrane potentials. We solely use the summed output potentials of the MSO and the lLSO and do not take into account, that in real biological systems the potentials would be converted into action potentials (APs). Since an AP is generated in an all-or-nothing manner, i.e., an AP is either elicited once a specific membrane potential threshold is reached or it is not elicited, AP generation is a highly non-linear transformation of a continuous summed input. To account for this non-linear transformation, at a very early stage of implementing the model, we introduced a non-linearity as described in Lehnert et al. (2014), where the firing rate, or rather firing probability, was dependent on input frequency and amplitude of inputs. The results showed to not be more beneficial than when dropping the non-linearity, i.e. there was no difference of the output except in the scaling of the calibration curve axes. Therefore we opted to drop the aspect of AP generation in favor of computation time.

3.2.4 *Only Considering the MSO*

A question that naturally arises is how do the results of the calibration curves change if we only consider one nucleus, namely the MSO since the lLSO's role in ITD estimation has long been unknown. In an early stage of model development, we extracted the calibration curves from the population membrane potentials of the MSO when comparing the left and right hemisphere with an input pure tone stimulus driven at $\mathrm{BF} = 200\mathrm{Hz}$. In this scenario an $\mathrm{ITD} = 0\mu s$ would result in ellipses collapsing on the midline and thus resulting in an angle of 45°. ITDs with opposite signs would always appear as ellipses mirrored at midline due to the symmetry of the ITD. The largest ITDs (absolute value) would scatter closest to x-axis and the y-axis. This resulted in calibration curves that were symmetrical to the y-axis and strictly monotonous s-shaped and their MPAs ranged from 12° to 78°. For higher BFs, all the ellipses would move closer to midline, resulting in smaller MPA intervals but all values symmetric around 45°. In general, the MPA ranges were very similar to the ones we can see for our analysis when considering both nuclei (see graphs in Figure 2.5). The major difference is of course, that the calibration curves are not strictly monotonous and not s-shaped anymore. To keep faithful to the underlying biology, we early on opted to consider both nuclei rather than just one.

3.2.5 *Input Signal Amplitude*

Another key parameter in our model is the amplitude A of the input signals. All input signals were normalized to correspond to $60\mathrm{dB\ SPL}$. This means that all sounds used to perform calibration have the exact same intensity or loudness. In nature, the loudness of sounds is of course not constant, but highly variable. To be biologically feasible, the proposed ITD encoding mechanism must be invariant towards amplitude modulation, because otherwise there would have to exist an infinite amount of sets of calibration curves. More precisely, the number of sets would be equal to the cardinality of $\mathbb{R}^+$ which is uncountably infinite.

To test for loudness invariance, we therefore calculated the calibration curves for varying amplitudes, namely $A' = A \times l$ for $l \in \{\frac{1}{4}, \frac{1}{2}, 1, 2, 3\}$. Interestingly, the resulting calibration curves were identical for all parameters l (Figure 3.1). This can be explained by the fact that the ITDs are encoded by the MPAs which are themselves retrieved from the 2d-membrane-potential-space of lLSO vs. MSO. If the input amplitude is A for MSO and lLSO, then the output membrane potential of MSO and lLSO will have amplitude A_m and A_l, respectively. If the input amplitude A is multiplied with a scaling factor $s \in \mathbb{R}$, then the new MSO and lLSO outputs will have amplitude $\hat{s} \times A_m$ and $\hat{s} \times A_l$, respectively, and $\hat{s}$ is the factor s after cochlear compression. It is evident, that this cannot change the calibration curves, as $\hat{s}$ is the identical factor for both nuclei and thus the MPA for the amplitude modulated signals remains the same when compared to the MPA which is retrieved for the signals without amplitude modulation.

We therefore conclude, that the here implemented ITD encoding mechanism is invariant under amplitude modulation. Furthermore, our implemented model also satisfies its sole intended purpose: It should only encode the information of the position in horizontal space (ITD) and not other features.

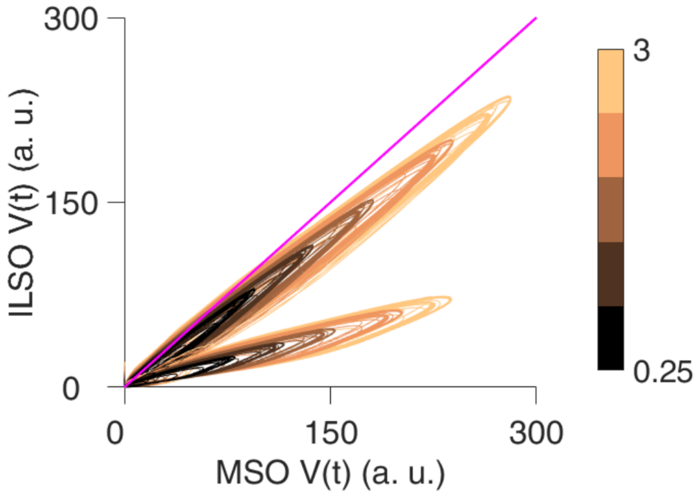

Figure 3.1: MPA amplitude invariance. Mean population responses of lLSO vs. MSO at BF = 200Hz for pure tone at two different fixes ITDs (top for 800µs and bottom for −300µs). Different colors represent different amplitude modulating factors $l \in \{\frac{1}{4}, \frac{1}{2}, 1, 2, 3\}$. Identity line in magenta for reference. The MPA is identical for all cases which proves invariance of the model to varying input signal amplitudes.

3.2.6 *Jitter*

Another parameter that we varied was latency jitter, i.e., the duration of transduction at the IHCs. Basis in our model is a flat distribution from $]0;1[\mathrm{ms}$. We investigated how the calibration curves would change when we would reduce the latency to (a) non-existent ($0\mathrm{ms}$) and (b) if we would extend the interval to $]0;2[\mathrm{ms}$. In both cases, there was no significant change noticeable. The ellipse-like objects as seen in Figure 2.4 only displayed a noticeable change in the length of the minor axis. For case (a) there was reduction and in case (b) there was an extension of its length as compared to the base settings. This change in length was symmetric to the principal axis. The more data points that are analyzed, the more insignificant a change of the minor axis becomes for the MPA. To recapitulate, the MPA is defined as the angle enclosed by the axis of abscissas and the population response vector $\vec{z}_{\mathrm{ITD}}$ (sum of all data points). For smaller neuron populations, the same latency jitters could impose a problem and result in significant changes for the MPA. Thus, a sufficient population size is needed to guarantee that the ITD encoding mechanism is robust towards the different jitter conditions that we tested for.

3.2.7 *Short-Term and Long-Term ITD Estimation Windows*

The short-term ITD estimation window was set to $\delta t = 10\mathrm{ms}$ in first analyses, i.e., every $10\mathrm{ms}$ an ITD estimation would be performed. However, it is apparent, that the width of the short-term ITD estimation window bounds the length of the long-term ITD estimation windows. For $10\mathrm{ms}$ this would mean, that for a long-term ITD of $100\mathrm{ms}$, this would result in only ten short-term ITDs. To this end, we shortened the short term ITD estimation window down to $1\mathrm{ms}$ with the underlying idea, that more short-term ITD estimations would make the long-term ITD estimation more stable by averaging out outliers (which have more weight when considering fewer estimates). The long-term ITD estimation window was found to be optimal if its length was greater or equal to $T = 1\mathrm{s}$. For this value, there was no strong deviance from the RMSE of estimated ITD and ground truth ITD over the pooled short-term ITDs from all observed BF-channels over the whole stimulus duration. We hypothesize a deviance of the long-term ITDs from the ground-truth ITDs for two main reasons. Firstly, we propose that the higher BF channels cause a distortion of good estimation, simply because the estimations are not very well performed in these channels. A remedy would therefore be to exempt these BF-channels from analysis or introduce some other weighting effect to mitigate their impact as it may be the case that these channels are not relevant for sound source localization. The second reason is probably due to the reduced availability of short-term ITDs for smaller T, as then the GMM might result in extreme over- or underfitting, causing a significant difference for estimated ITD and ground truth ITD. However, by manipulating the GMMs to gauge as many ITD estimators as signals were actually present, we found the RMSE to vanish for large windows T. Furthermore, when the GMM was made to gauge the numbers of signals, there was poor performance in areas of low signal overlap. From this we conclude that there might be two separate successive tasks to be able to perform sound localization. The first one is, that the amount of signals within a complex signal must first be extracted. This could always be done successfully by the GMMs. Once the correct number of signals has been found, then the precise localization of the sounds can be carried out. It can be argued that sound source localization improves by the integration of multimodal information, i.e., knowing how many sound sources are present (e.g., from visual input) aids the localization process (Suied et al., 2008). Furthermore, it is interesting to mention, that lateralization always worked within the model. This means, that the general direction of the sounds can always be determined, hinting at the fact, that maybe absolute sound localization is not as important as being able to separate sounds from each other in order to gain a quick general overview of the current auditory scene.

3.3 ENCODING SOUND SOURCE POSITIONS FROM TWO NUCLEI SIMULTANESOUSLY

In order to explain why we propose a population coding model of sound localization in mammals that makes use of both nuclei simultaneously, we now turn to the biological fundamentals of the fast and temporally precise processing of ITDs in the

MSO and LSO. For this, it is necessary to understand their evolutionary emergence, how their integration process works and why it herefrom follows, that the MSO and LSO outputs together can provide a reliable estimation of azimuthal sound source position.

3.3.1 *The Evolutionary "Choice" for ILDs in Early Mammals*

To be able to hear and localize airborne sounds, a device to perform impedance-matching is needed that can sufficiently vibrate. About 210 to 230 million years ago in the Triassic, the tympanum and the specialized middle ear evolved three times independently in amphibians (*Anura*), sauropsids and mammals due to common selection pressure (Allin, 1975; Clack, 1997). These two devices fulfill the prerequisites of detecting airborne sounds. The two main cues for sound localization are the ITD for low-frequency and the ILD for high-frequency sounds (cf. Section 1.2). But both cues could most likely not be exploited by early ancestors of mammals. For example, *Morganucodon* were smaller in size than mice. Even if they did use ITDs, their small head size could only account for ITDs up to 50µs. It is more likely that they mainly exploited ILDs in order to guarantee their survival and that there was simply no need to use ITDs. A small larynx could have produced sounds in the high-frequency range and their small head size could sufficiently make use of ILDs (Grothe and Pecka, 2014). As ILDs occur when the wavelength is shorter than the head, even small animals are able to detect significant ILDs at high frequencies (Erulkar, 1972; Harnischfeger et al., 1985). Most importantly, high-frequency communication avoids predation through birds as they are low-frequency hearers below 10kHz (Grothe, 2003; Grothe and Pecka, 2014). All this points to the assumption, that early mammals only used ILDs to locate sounds.

Fast Temporal Processing in the LSO.

The LSO is the nucleus which is sensitive to ILDs (Galambos et al., 1959; Boudreau und Tsuchitani, 1968) and the integration process is a summation of one ipsilateral excitatory and one contralateral inhibitory input (see Section 1.1.4). Therefore LSO cells are ipsileading, because an ipsilateral positioned sound ($\mathrm{dB} < 0$) will create a strong excitation and the contralateral ($\mathrm{dB} > 0$) inhibition will be much lower due to the sound shadow at the head. The tuning curves of LSO neurons are therefore sigmoids with steep slopes at midline ($\mathrm{dB} = 0$), low activity levels for positive ILDs and high activity levels for negative ILDs (Tollin, 2003). As ILD detectors, the LSO neurons capture the timing of fluctuations in signal amplitudes. To be able to achieve this, there are two main requirements, namely short integration times and a coincidence detection mechanism at LSO stage.

Short Integration Times. To be able to be detect amplitude fluctuations in a temporally precise manner, LSO cells must have short integration times. A mechanism that integrates and averages over whole stimulus durations is extremely unlikely to capture fast timed events, because the average could theoretically change during the integration process, making it impossible to localize more than one sound source. For complex sounds, which contain many sharp rising amplitude modulations (transients), very short integration windows are necessary (Grothe and Pecka, 2014; Grothe et al., 2019). Transients in complex signals are encoded in the spiral ganglion/auditory nerve via phase locking and maintained in the afferent circuitry and then used by LSO cells which guarantees high fidelity of timing information. After the precise timing of auditory events is captured by phase locking in the AN, this timing information is conserved by the SBCs and GBCs and transferred in high fidelity to their next respective stages (Joris et al., 1994). The GBCs (contralateral input) project onto cells of the MNTB via the largest synapses of the brain, the Calyx of Held. This sizeable synapse is well studied for its fast and high fidelity transmission (Schneggenburger and Forsythe, 2006; Kopp-Scheinpflug et al., 2011). It is not only extremely fast due to low synaptic latency, it is also very reliable in that a presynaptic AP will (almost) always result in a postsynaptic AP and lastly has low jitter which also helps maintain phase locking. The input of the MNTB is then projected onto the LSO. The SBCs (ipsilateral input) project directly onto the LSO. Phase locking not only occurs for sounds at low frequencies as discussed in Section 1.1.3, but at sounds of all frequencies under the condition that the signal contains transients or onsets (Dietz et al., 2014). The membrane kinetics of LSO cells are able to preserve the high temporal precision of the MNTB and the

SBC input (Tollin, 2003). In short, the expression and complex interaction of two specific channels, namely fast opening low-threshold Kv1 channels (Mathews et al., 2010) and slow opening HCN channels (Baumann et al., 2013), results in membrane time constants of about $\tau \sim 0.5 - 1.5ms$ (Wu and Kelly, 1991; Sanes and Takács, 1993; Gittelman and Tempel, 2006; Pilati et al., 2016) which results in short integration time windows. For a membrane time constant of $\tau = 0.5ms$, this would translate into an upper bound of the phase locking of the SBCs and GBCs at $2kHz$. Note, that for any phase-locked input with the same or a lower firing rate, the LSO kinetics guarantee timing sensitivity on a cycle-by-cycle basis (Grothe et al., 2019). All of these specializations and components of the LSO are thus responsible for high sensitivity of timing information, which corroborates our theory, that the LSO could potentially be exploited – together with the MSO – for ITD encoding.

Coincidence Detection in the LSO. Another requirement to achieve temporally precise ILD processing, is not only that the information is preserved from the stage of the AN until it reaches the LSO, but also that the excitatory and inhibitory inputs must occur in close proximity to each other or else the inputs could not influence one another, i.e., no subtraction or addition processes of the EPSPs (excitatory postsynaptic potentials) and IPSPs (inhibitory postsynaptic potentials) during integration could take place (Park et al., 1996; Tollin, 2003; Grothe and Pecka, 2014). Joris and Yin (2005) illustrate this problem in the case for a sound coming straight from ahead which would correspond to $0dB$ ILD (and also vanishing ITD). In order for the excitation and inhibition to coincide at the LSO, there are two main problems that have to be overcome. Firstly, the pathway to the LSO from both ears is not equidistant, i.e., the contralateral path is longer than the ipsilateral path. Secondly, the contralateral pathway has an additional synaptic relay stage, the Calyx of Held synapse (Tollin, 2003). In order to be able to counter the anatomical differences in both pathways, several specializations have been reported. To compensate for the the overall differences in pathway length, the axons from the GBCs leading to the MNTB have a greater diameter (Morest, 1968) and thicker myelination (Ford et al., 2015) as compared to the axons of the SBCs. Both morphological features that increase velocity of conduction (Grothe and Pecka, 2014). The Calyx of Held guarantees incredibly fast synaptic transmission due to hundreds of active zones and a large vesicle pool in the presynaptic bouton (Taschenberger et al., 2002; Yu and Goodrich, 2014), resulting in one of the shortest synaptic delays ever measured in the central nervous system (CNS). These specializations guarantee precise coincidence detection at the LSO.

Taking all results together, the head size of small ancestral mammals was probably the reason for the evolutionary "choice" of utilizing ILDs for sound localization. The possibility of processing ILDs came along with biophysical, anatomical and physiological challenges that had to be overcome in order to retain the temporal information of input sounds at the stage of the LSO. These challenges were overcome by specific evolutionary adaptations within the ILD circuit resulting in a coincidence detector mechanism with integration time windows so short, that they can retain the phase-locked temporal information of the input sounds. Most importantly for us, the short integration times and the coincidence detection mechanism of the LSO make it reasonable to include this nucleus in our model of ITD estimation, rather than restricting the LSO's functional use solely to ILDs, since all of the described mechanisms and specializations guarantee high fidelity of timing information.

3.3.2 *Combining Information from the MSO and LSO.*

During evolution, when mammals increased in size, not only their head and larynx became larger, but they also inhabited larger territories which required communication over long distances using low-frequency sounds which travel farther than high-frequency sounds before their energy dissipates (Grothe and Pecka, 2014). These animals were already sensitive to ILDs by possessing a well-developed LSO. But low-frequency sounds produce dismissible small ILDs so the need for a precise ITD extraction mechanism grew. This was probably the driving selection pressure for adaptations in the MSO to be sensitive to ITDs. That the MSO is most relevant for sound localization in mammals via ITDs seems to be the scientific consensus nowadays (Spitzer and Semple, 1995; Pecka et al., 2008; Franken, Bremen and Joris, 2014), so we shall not delineate the properties[3] that explain how it achieves such high temporal precision, but rather discuss why it is reasonable to assume, that the human binaural system makes use of both nuclei – MSO and LSO – simultaneously for ITD estimations. The LSO, as described in the previous section, was already specialized for retaining temporal features. In cat and chinchilla (Joris and Yin, 1995; Tollin and Yin, 2002), when presented with ITDs on a microsecond scale, there were strong response rate modulations measured in their lLSO neurons, which shows that the neurons of the low-frequency limb of the LSO are capable of cycle-by-cycle ITD sensitivity via phase locking. The MSO is seen as a refined LSO (Grothe and Pecka, 2014). The refinement lies within the doubling of the pathways already present in the LSO. The MSO has not only two pathways leading it, but two additional ones, i.e., the MSO receives two excitatory and two inhibitory inputs. The different amount of inputs to the MSO results in a different ITD tuning curve as compared to the LSO (see Figure 1.8). The LSO neurons are ipsileading (what one would expect for an excitatory-inhibitory coincidence detector, where excitation precedes inhibition), the MSO neurons are contraleading (which is likely due to the complex interplay of the four inputs and the special role of fast glycinergic inhibition preceding excitation, cf. Roberts et al., 2013 and Myoga et al., 2014). Since the hemispheric preference of the nuclei is switched for the LSO at higher stages, but stays the same for the MSO, their individual population codes are compatible at the midbrain/cortex stage, e.g., the higher the overall activity is in one hemisphere, the more contralateral the sound source would be located, and vice versa.

Concluding, there are two main similarities between the MSO and LSO. Both nuclei act as coincidence detectors, i.e., they have the same design principle. Furthermore, the integration mechanism of both coincidence detectors results in a population code of ITD that each cover a wide range of acoustic space. Both nuclei thus share the same coding principle. Because of these shared principles and under the premise that both nuclei (within one hemisphere) should make the same estimation for a sound at a specific location in space, it is therefore not unreasonable to make the assumption that this position can be read out by evaluating the output of MSO and LSO simultaneously, establishing a 2d-code auf auditory space.

[3] In short, some of the most important properties are that the MSO cells have membrane time constants of only ~ 300μs and morphological adaptations to increase conduction velocity (similar to the LSO). For a review on further properties see Grothe et al. (2019)

3.4 CONCLUDING REMARKS

We have presented a novel model of sound source coding in the azimuthal plane which is biologically motivated. Different selection pressures and different biological prerequisites have successively brought forward two separate binaural coincidence detection systems, the LSO for ILDs and the MSO for ITDs. The capability of the LSO to also process ITDs as well as the the circuit and coding properties shared with the MSO led us to believe, that the interaction between the two nuclei is necessary to obtain a clear mapping of auditory space. We then presented a method of sound source separation solely based on one cue, the ITD, which was obtained from the responses of both nuclei. The motivation for analyzing sound separation as the objective of the two population codes is due to the following evolutionary assumptions. Consider a sleeping animal, which is awaken by a growl. It could run to safety if it correctly localizes the position of the predator. At first look, to be able to know the absolute position of the predator thus seems to be the straightforward objective of sound localization. But this perspective misses an important point: Mammals are animals that are usually not stationary or fixed to one point in space. Mammals move actively around in the world. They prey on other animals or are being preyed on themselves. Their auditory surroundings are subject to great variation. Many of the earliest mammals, such as *Morganucodon*, lived in a world of darkness, having adapted to a nocturnal life while living in coexistence with the roaming dinosaurs (Grothe and Pecka, 2014). They lived in a world where constant alertness and having to be in motion could be crucial for survival. The fast movement would of course also change all of the absolute positions of all sound sources in a matter of seconds. In short, auditory scenes are highly dynamic and context-dependent. Therefore, the idea of being able to successfully separate sound sources in such a complex scene seems to be a plausible evolutionary driving force for communication and survival. However, the same evolutionary pressure might not have been the same for all animals. Whereas for smaller animals sound separation might have been of paramount importance for communication and survival in order to avoid predation, for larger animals, sound localization acuity might have been more important in order to correctly localize its prey and thus guarantuee their survival. Therefore, it can only be speculated about the true objective of sound localization.

Furthermore it has been shown, that the tuning curves of the MSO and LSO are not fixed, but their output is modulated by GABAergic feedback (Stange et al., 2013). Activation of the nuclei leads to presynaptic inhibition by activation of GABA-B receptors, which is mediated by an additional nucleus, the superior paraolivary nucleus (SPN), and results in a dampening of the amplitude of the resulting output (i.e., gain regulation through negative feedback). This results in a shift of the perceived midline (zero azimuth) towards where the sound was coming from and thus would also shift the perceived location of a subsequent sound. Most recently, in a conductance-based model of LSO neurons investigating the encoding of ILDs, it has been shown, that the adaptation introduced by the GABAergic feedback loop results in a shift of coding precision towards the ILD of the adapter tone which results in a higher sensitivity for ILDs in adapter proximity (Oess et al., 2020). The effect of a shifted perceived midline after adaptation has also been documented in psychophysical experiment (Lingner et al., 2018). Since this inhibitory feedback loop is found in both the MSO and LSO circuit, it is assumed, that the main evolutionary constraint to shape the nuclei into its today's form was not necessarily absolute sound localization, but sound separation, which marks a paradigm shift in auditory neuroscience as to what the true underlying principle of sound localization may be.

A final question we might ask at this point is how can the findings of this thesis be embedded in the historical and future scientific landscape. From the historical viewpoint it is striking, that when Durlach proposed his *Equalization and Cancellation Model* in 1963, his findings from over half a century ago have hinted at the relevance of not exact absolute sound source localization acuity, but rather the importance of relative sound source separation. His underlying mechanism using a subtraction operation acting on both ears showed, that a previously occurring sound (the masker) can help improve ITD discrimination; similar to what Lingner et al. (2018) and Oess et al. (2020) have suggested quite recently. Even though our effective model makes use of a completely different ITD-decoding approach (hemispheric balancing), it could still be tested for the relevance of relative sound source separation. The effective model presented in this thesis is rather powerful for providing stream separation and we have learned, that the ITD can theoretically serve as the sole cue for sound source localization.

In order to move forward with this model and to validate it in future experiments, the effective model should be expanded to incorporate the aforementioned GABAergic feedback loop and activity-dependent input modulation. In addition, it should also be investigated how well the ITD serves as a sound separating cue under the condition, that the sounds are not presented concurrently (both starting at $t = 0$), but one after another (i.e., sound one at $t_1 = 0$, sound two at $t_2 \neq t_1$). This would serve as a test of the effective model to see if it can reproduce the results as observed in Lingner et al. (2018) and Oess et al. (2020). Until the model is not adapted to reflect all known aspects of the true underlying biology, we cannot infer if the presented decoding mechanisms resemble some or any of the functionality of the brain circuits. Even in the case that the feedback loop were to be incorporated, it would still be highly speculative to link the presented decoding mechanisms to any specific part of the circuits. This is due to the nature of the design of the effective model's most important quantity, the MPA. It simply cannot be directly connected to a specific structure in the auditory brainstem. This is owing to the fact, that the three defining properties BF, CD and CP of an MSO and lLSO neuron can by interpreted as a bypass or a substitute for the entire neuronal circuit in the auditory brainstem, i.e., we only make use of the information that arrives at the neurons directly and not at the stages in-between ears and neurons. Nevertheless, the MPA still serves as a powerful proof of principle, that a potential sound source separating mechanism of humans could completely be based on ITD.

Finally, even though we have found a working model of sound source separation based on ITD, it only seems to represent one of many possible answers to a question that might be the wrong one to ask in the first place. The irrefutable truth is that no coding theory can conclude on what the true underlying objective function is, that the code is trying to satisfy (Leibold and Grothe, 2015). We can only guess. Nevertheless, I hope that this research can contribute to a better understanding of sound localization theory. Questions were answered, but many questions remain open.

4

REFERENCES

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle." In: Breakthroughs in Statistics I. *Springer*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Allin, E. F. (1975). Evolution of the mammalian middle ear. *Journal of Morphology*, 147(4), 403–437.

Baumann, V. J., Lehnert, S., Leibold, C., and Koch, U. (2013). Tonotopic organization of the hyperpolarization-activated current (Ih) in the mammalian medial superior olive. *Frontiers in Neural Circuits*, 7, 117.

Bear, M. F., Connors, B. W., and Paradiso, M. A. (2016). Neuroscience: exploring the brain. Fourth edition. *Wolters Kluwer*.

Beckius, G. E., Batra, R. and Oliver, D. L. (1999). Axons from anteroventral cochlear nucleus that terminate in medial superior olive of cat: observations related to delay lines. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 19, 3146–3161.

Blauert, J. (1997). Spatial hearing: the psychophysics of human sound localization. *MIT Press*.

Boudreau, J. C., and Tsuchitani, C. (1968). Binaural interaction in the cat superior olive S segment. *Journal of Neurophysiology*, 31(3), 442–454.

Bray, D. (2015). Limits of computational biology. *In Silico Biology*, 12(1,2), 1–7.

Breebaart, J., van de Par, S., and Kohlrausch, A. (2001). Binaural processing model based on contralateral inhibition. I. Model structure. *The Journal of the Acoustical Society of America*, 110(2), 1074–1088.

Brown, M. C., and Santos-Sacchi, J. (2012). "Audition." In: Fundamental neuroscience. Fourth edition. *Academic Press*.

Carr, C. E., and Konishi, M. (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl. *The Journal of Neuroscience*, 10(10), 3227–3246.

Colburn, H. S. (1973). Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination. *The Journal of the Acoustical Society of America*, 54(6), 1458–1470.

Colburn, H. S. (1977). Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise. *The Journal of the Acoustical Society of America*, 61(2), 525–533.

Colburn, H. S., and Durlach, N. I. (1978). "Models of binaural interation." In: Handbook of perception. *Academic Press*.

Colburn, H. S., and Kulkarni, A. (2005). "Models of sound localization." In: Sound source localization (Springer handbook of auditory research). *Springer*.

Cooke, M. (2003). Glimpsing speech. *Journal of Phonetics*, 31(3-4), 579–584.

Clack, J. A. (1997). The evolution of tetrapod ears and the fossil record. *Brain, Behavior and Evolution*, 50(4), 198–212.

Day, M. L., and Delgutte, B. (2013). Decoding Sound Source Location and Separation Using Neural Population Activity Patterns. *Journal of Neuroscience*, 33(40), 15837–15847.

Day, M. L., and Semple, M. N. (2011). Frequency-dependent interaural delays in the medial superior olive: implications for interaural cochlear delays. *Journal of Neurophysiology*, 106(4), 1985–1999.

Dietz, M., Ewert, S. D., Hohmann, V., and Kollmeier, B. (2008). Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences. *Brain Research*, 1220, 234–245.

Dietz, M., Marquardt, T., Stange, A., Pecka, M., Grothe, B., and McAlpine, D. (2014). Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds II: single-neuron recordings. *Journal of Neurophysiology*, 111(10), 1973–1985.

Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *The Journal of the Acoustical Society of America*, 35(8), 1206–1218.

Encke, J. and Hemmert, W. (2018). Extraction of inter-aural time differences using a spiking neuron network model of the medial superior olive. *Frontiers in Neuroscience*, 12, 140–141.

Erulkar, S. D. (1972). Comparative aspects of spatial localization of sound. *Physiological Reviews*, 52(1), 237–360.

Fischer, L., Leibold, C., and Felmy, F. (2018). Resonance properties in auditory brainstem neurons. *Frontiers in Cellular Neuroscience*, 12.

Ford, M. C., Alexandrova, O., Cossell, L., Stange-Marten, A., Sinclair, J., Kopp-Scheinpflug, C., Pecka, M., Attwell, D., Grothe, B. (2015). Tuning of Ranvier node and internode properties in myelinated axons to adjust action potential timing. *Nature Communications*, 6(1). 8073.

Franken, T. P., Bremen, P., and Joris, P. X. (2014). Coincidence detection in the medial superior olive: mechanistic implications of an analysis of input spiking patterns. *Frontiers in Neural Circuits*, 8, 42.

Franken, T. P., Roberts, M. T., Wei, L., Golding, N. L., and Joris, P. X. (2015). In vivo coincidence detection in mammalian sound localization generates phase delays. *Nature Neuroscience*, 18(3), 444–452.

Galambos, R., and Davis, H. (1943). The response of single auditory-nerve fibers to acoustic stimulation. *Journal of Neurophysiology*, 6(1), 39–57.

Galambos, R., Schwartzkopff, J., and Rupert, A. (1959). Microelectrode study of superior olivary nuclei. *American Journal of Physiology-Legacy Content*, 197(3), 527–536.

Gittelman, J. X., and Tempel, B. L. (2006). Kv1.1-containing channels are critical for temporal precision during spike initiation. *Journal of Neurophysiology*, 96(3), 1203–1214.

Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47, 103-138.

Goodman D. F. M., Benichoux V., and Brette R. (2013) Decoding neural responses to temporal cues for sound localization. *Elife*, 2:e01312.

Goldberg, J. M., and Brown, P. B. (1969). Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. *Journal of Neurophysiology*, 32(4), 613–636.

Groß, S., and Leibold, C. (2019). Short-term ITD (interaural time difference) estimation of natural sound stimuli via effective models of binaural brainstem nuclei. [Abstract and poster presentation at conference *Neurowissenschaftliche Gesellschaft Göttingen*].

Groß, S., and Leibold, C. (2020). Sound source location in the azimuthal plane: separating sounds via short-term interaural time difference estimations. [Abstract and poster presentation at conference *Bernstein Conference Berlin*].

Grothe, B. (2003). New roles for synaptic inhibition in sound localization. *Nature Reviews Neuroscience*, 4(7), 540–550.

Grothe, B., and Pecka, M. (2014). The natural history of sound localization in mammals – a story of neuronal inhibition. *Frontiers in Neural Circuits*, 8.

Grothe, B., Pecka, M., and McAlpine, D. (2010). Mechanisms of sound localization in mammals. *Physiological Reviews*, 90(3), 983–1012.

Grothe, B., Leibold, C., and Pecka, M. (2019). "The medial superior olivary nucleus: meeting the need for speed." In: The Oxford handbook of the auditory brainstem. *Oxford University Press*.

Hancock, K. E. and Delgutte, B. (2004). A physiologically based model of interaural time difference discrimination. *Journal of Neuroscience*, 24(32), 7110–7117.

Hancock, K. E. (2006). "A physiologically-based population rate code for interaural time differences (ITDs) predicts bandwidth-dependent lateralization." In: Hearing – From Sensory Processing to Perception. *Springer*, 389-397.

Hall, J. L. (1964). Binaural interaction in the accessory superior olivary nucleus of the cat - an electrophysiological study of single neurons. [Technical report 416; [...] for partial fullfilment of the requirements for the degree of Doctor of Philosophy, Massachusetts Institute of Technology]. dspace.mit.edu/handle/1721.1/4409.

Harper, N. S., and McAlpine, D. (2004). Optimal neural population coding of an auditory spatial cue. *Nature*, 430(7000), 682–686.

Harnischfeger, G., Neuweiler, G., and Schlegel, P. (1985). Interaural time and intensity coding in superior olivary complex and inferior colliculus of the echolocating bat Molossus ater. *Journal of Neurophysiology*, 53, 89–109.

Hilbig, H., Beil, B., Hilbig, H., Call, J., and Bidmon, H.-J. (2009). Superior olivary complex organization and cytoarchitecture may be correlated with function and catarrhine primate phylogeny. *Brain Structure and Function*, 213(4-5), 489–497.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41(1), 35–39.

Joris, P. X., Carney, L. H., Smith, P. H., and Yin, T. C. (1994). Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency. *Journal of Neurophysiology*, 71(3), 1022–1036. doi:10.1152/jn.1994.71.3.1022

Joris, P. X., and Yin, T. C. (1995). Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences. *Journal of Neurophysiology*, 73(3), 1043–1062.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., and Hudspeth, A. J. (2013) Principles of neural science. Fifth edition. *The McGraw-Hill Companies*.

Karino, S., Smith P. H., Yin, T. C., and Joris, P. X. (2011). Axonal branching patterns as sources of delay in the mammalian auditory brainstem: a re-examination. Journal of Neuroscience, 31, 3016–3031.

Klug, J., Schmors, L., Ashida, G. and Dietz, M. (2020). Neural rate difference model can account for lateralization of high-frequency stimuli. *The Journal of the Acoustical Society of America*, 148(2), 678–691.

Klumpp, R. G., and Eady, H. R. (1956). Some Measurements of Interaural Time Difference Thresholds. *The Journal of the Acoustical Society of America*, 28(5), 859–860.

Kopp-Scheinpflug, C., Steinert, J. R., and Forsythe, I. D. (2011). Modulation and control of synaptic transmission across the MNTB. *Hearing Research*, 279(1-2), 22–31.

Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, 62(1), 157–167.

Lehnert, S., Ford, M. C., Alexandrova, O., Hellmundt, F., Felmy, F., Grothe, B., and Leibold, C. (2014). Action potential generation in an anatomically constrained model of medial superior olive axons. *Journal of Neuroscience*, 34(15), 5370–5384.

Lehnert, S. (2015). Biophysical principles underlying binaural coincidence detection: Computational approaches. [Doctoral dissertation, Ludwig-Maximilians-Universität München]. edoc.ub.uni-muenchen.de/19033/1/Lehnert_Simon.pdf.

Leibold, C., and Grothe, B. (2015). Sound localization with microsecond precision in mammals: what is it we do not understand? *e-Neuroforum*, 6(1), 3–10.

Lindemann, W. (1986). Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *The Journal of the Acoustical Society of America*, 80(6), 1608–1622.

Lingner, A., Pecka, M., Leibold, C., and Grothe, B. (2018). A novel concept for dynamic adjustment of auditory space. *Scientific Reports*, 8(1).

Lüling, H., Siveke, I., Grothe, B., and Leibold, C. (2011). Frequency-invariant representation of interaural time differences in mammals. *PLoS Computational Biology*, 7(3), e1002013.

Luo, L. (2015) Principles of neurobiology. *Garland Science*.

Maki K, and Furukawa, S. (2005). Acoustical cues for sound localization by the Mongolian gerbil, Meriones unguiculatus. *The Journal of the Acoustical Society of America*, 118, 872-886.

Mathews, P. J., Jercog, P. E., Rinzel, J., Scott, L. L., and Golding, N. L. (2010). Control of submillisecond synaptic timing in binaural coincidence detectors by Kv1 channels. *Nature Neuroscience*, 13(5), 601–609.

McAlpine, D., Jiang, D., and Palmer, A. R. (2001). A neural code for low-frequency sound localization in mammals. *Nature Neuroscience*, 4(4), 396–401.

McFadden, D., and Pasanen, E. G. (1976). Lateralization at high frequencies based on interaural time differences. *The Journal of the Acoustical Society of America*, 59(3), 634–639.

Meddis, R., and Lopez-Poveda, E. A. (2010). "Auditory periphery: from pinna to auditory nerve". In: Computational models of the auditory system (Springer handbook of auditory research). *Springer*.

Mi, J., Groll, M., and Colburn, H. S. (2017). Comparison of a target-equalization-cancellation approach and a localization approach to source separation. *The Journal of the Acoustical Society of America*, 142(5), 2933–2941.

Mills, A. W. (1958). On the Minimum Audible Angle. *The Journal of the Acoustical Society of America*, 30(4), 237–246.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26(9), 394–95.

Morest, D. K. (1968). The collateral system of the medial nucleus of the trapezoid body of the cat, its neuronal architecture and relation to the olivo-cochlear bundle. *Brain Research*, 9(2), 288–311.

Myoga, M. H., Lehnert, S., Leibold, C., Felmy, F., and Grothe, B. (2014). Glycinergic inhibition tunes coincidence detection in the auditory brainstem. *Nature Communications*, 5(1), 3790.

Oess, T., Ernst, M. O., Neumann, H. (2020). Computational principles of neural adaptation for binaural signal integration. PLOS Computational Biology, 16(7), e1008020–.

O'Reilly, R. C., and Munakata, Y. (2000). Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain. *The MIT Press*.

Oxenham, A. J., and Moore, B. C. J. (1994). Modeling the additivity of nonsimultaneous masking. Hearing Research, 80(1), 105–118.

Park, T. J., Grothe, B., Pollak, G. D., Schuller, G., Koch, U. (1996). Neural delays shape selectivity to interaural intensity differences in the lateral superior olive. *The Journal of Neuroscience*, 16(20), 6554–6566.

Patterson, R. D. and Moore, B. C. J. (1986). "Auditory filters and excitation patterns as representations of frequency resolution". In: Frequency Selectivity in Hearing. *Academic Press*.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand M. (1992). "Complex sounds and auditory images". In: Auditory physiology and perception. *Pergamon*.

Pecka, M., Brand, A., Behrend, O., and Grothe, B. (2008). Interaural time difference processing in the mammalian medial superior olive: the role of glycinergic Inhibition. *Journal of Neuroscience*, 28(27), 6914–6925.

Penrose, R. (1955). A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51(3), 406–13.

Pilati, N., Linley, D. M., Selvaskandan, H., Uchitel, O., Hennig, M. H., Kopp-Scheinpflug, C., and Forsythe, I. D. (2016). Acoustic trauma slows AMPA receptor-mediated EPSCs in the auditory brainstem, reducing GluA4 subunit expression as a mechanism to rescue binaural function. *The Journal of Physiology*, 594(13), 3683–3703.

Remme, M. W. H., Donato, R., Mikiel-Hunter, J., Ballestero, J. A., Foster, S., Rinzel, J. and McAlpine, D. (2014). Subthreshold resonance properties contribute to the efficient coding of auditory spatial cues. *Proceedings of the National Academy of Sciences*, 111(22), E2339–E2348.

Rhode, W. S., and Smith, P. H. (1986). Encoding timing and intensity in the ventral cochlear nucleus of the cat. *Journal of Neurophysiology*, 56(2), 261–286.

Roberts, M. T.; Seeman, S. C.; Golding, N. L. (2013). A mechanistic understanding of the role of feedforward inhibition in the mammalian sound localization circuitry. *Neuron*, 78(5), 923–935.

Sanes, D. H., and Takács, C. (1993). Activity-dependent refinement of inhibitory connections. *European Journal of Neuroscience*, 5(6), 570–574.

Schneggenburger, R., and Forsythe, I. D. (2006). The calyx of Held. *Cell and Tissue Research*, 326(2), 311–337.

Schnupp, J., Nelken, I., and King, A. (2011). Auditory neuroscience: Making sense of sound. *MIT Press*.

Schwarz, G. E. (1978). Estimating the dimension of a model. Annals of Statistics, 6(2), 461–464.

Seidl, A. H., Rubel, E. W., and Harris, D. M. (2010). Mechanisms for adjusting interaural time differences to achieve binaural coincidence detection. *Journal of Neuroscience*, 30(1), 70–80.

Siveke, I., Pecka, M., Seidl, A. H., Baudoux, S., and Grothe, B. (2006). Binaural response properties of low-frequency neurons in the gerbil dorsal nucleus of the lateral lemniscus. *Journal of Neurophysiology*, 96(3), 1425–1440.

Siveke, I., Leibold, C., and Grothe, B. (2007). Spectral composition of concurrent noise affects neuronal sensitivity to interaural time differences of tones in the dorsal nucleus of the lateral lemniscus. *Journal of Neurophysiology*, 98(5), 2705–2715.

Spitzer, M. W., and Semple, M. N. (1995). Neurons sensitive to interaural phase disparity in gerbil superior olive: diverse monaural and temporal response properties. *Journal of Neurophysiology*, 73(4), 1668–1690.

Stange A., Myoga, M. H., Lingner, A., Ford, M.C., Alexandrova, O., Felmy, F., Pecka, M., Siveke, I., Grothe, B. (2013) Adaptation in sound localization: from GABA(B) receptor-mediated synaptic modulation to perception. *Nature Neuroscience*, 16, 1840–1847.

Stecker, G. C., and Gallun, F. J. (2012). "Binaural hearing, sound localization, and spatial hearing". In: Translational perspectives in auditory neuroscience: normal aspects of hearing. *Plural Publishing*.

Stern, R. M., and Colburn, H. S. (1978). Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position. *The Journal of the Acoustical Society of America*, 64(1), 127–140.

Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (1988). Lateralization of complex binaural stimuli: A weighted-image model. *The Journal of the Acoustical Society of America*, 84(1), 156–165.

Strutt J. W. (Lord Rayleigh) (1907) XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13:74, 214-232.

"Sound." (2011). In: *merriam-mebster.com*. Retrieved May 16, 2021, from https://www.merriam-webster.com/dictionary/sound.

Suied, C., Bonneel, N., and Viaud-Delmon, I. (2008). Integration of auditory and visual information in the recognition of realistic objects. *Experimental Brain Research*, 194(1), 91–102.

Takanen, M., Santala, O. and Pulkki, V. (2014). Visualization of functional count-comparison-based binaural auditory model output. *Hearing Research*, 309, 147–163.

Taschenberger, H., Leo, R. M., Rowland, K. C., Spirou, G. A., and von Gersdorff, H. (2002). Optimizing synaptic architecture and efficiency for high-frequency transmission. *Neuron*, 36(6), 1127–1143.

Teufel, C., and Fletcher, P. C. (2016). The promises and pitfalls of applying computational models to neurological and psychiatric disorders. *Brain*, 139(10), 2600–2608.

Thavam, S., and Dietz, M. (2019). Smallest perceivable interaural time differences. The Journal of the Acoustical Society of America, 145(1), 458–468.

Tollin, D. J. (2003). The lateral superior olive: a functional role in sound source localization. *The Neuroscientist*, 9(2), 127–143.

Tollin, D. J., and Yin, T. C. T. (2005). Interaural phase and level difference sensitivity in low-frequency neurons in the lateral superior olive. *Journal of Neuroscience*, 25(46), 10648–10657.

van Bergeijk, W. A. (1962). Variation on a theme of Békésy: a model of binaural interaction. *The Journal of the Acoustical Society of America*, 34(9B), 1431–1437.

von Békésy, G. (1930). Zur Theorie des Hörens; Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkenungleicheit der beiderseitigen Schalleinwirkungen. *Physikalische Zeitschrift*, 31:824-35, 857-68.

von Békésy, G., and Wever, E. G. (1960). Experiments in hearing. *McGraw-Hill*.

Wit, E., van den Heuvel, E., Romeyn, J.-W. (2012). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3), 217–236.

Wu, S. H., and Kelly, J. B. (1991). Physiological properties of neurons in the mouse superior olive: membrane characteristics and postsynaptic responses studied in vitro. *Journal of Neurophysiology*, 65(2), 230–246.

Yin, T. C., and Kuwada, S. (1983a). Binaural interaction in low-frequency neurons in inferior colliculus of the cat. II. Effects of changing rate and direction of interaural phase. *Journal of Neurophysiology*, 50(4), 1000–1019.

Yin, T. C., and Kuwada, S. (1983b). Binaural interaction in low-frequency neurons in inferior colliculus of the cat. III. Effects of changing frequency. *Journal of Neurophysiology*, 50(4), 1020–1042.

Yin, T. C., and Chan, J. C. (1990). Interaural time sensitivity in medial superior olive of cat. *Journal of Neurophysiology*, 64(2), 465–488.

Yin, T. C. T., Smith, P. H., and Joris, P. X. (2019). Neural Mechanisms of Binaural Processing in the Auditory Brainstem. *Comprehensive Physiology*, 9(4), 1503–1575.

Yu, W.-M., and Goodrich, L. V. (2014). Morphological and physiological development of auditory synapses. *Hearing Research*, 311, 3–16.

Zenner, H.-P. (1994). Hören: Physiologie, Biochemie, Zell- und Neurobiologie. First edition. *Georg Thieme Verlag*.

Zwislocki, J., and Feldman, R. S. (1956). Just Noticeable Differences in Dichotic Phase. *The Journal of the Acoustical Society of America*, 28(5), 860–864.

ACKNOWLEDGEMENTS

I think I'll stop here.

— Andrew Wiles

The past six years have been quite the extraordinary journey. I have had the great opportunity to work on this thesis and it has been one of the most valuable and life-changing chapters in my life. This all would have not been quite the same experience without the help of several people that accompanied me during this time.

First and foremost, I would like to thank my supervisor Christian Leibold who gave an aspiring high school teacher the chance to learn about the compelling and complex field of neuroscience. If it had not been for him, I would have probably never even applied to a graduate school let alone a PhD position in this intricate area of science and I am very grateful to him that he encouraged me to do so and thereby unlocking my scientific potential. Christian always provided me not only with invaluable independence and freedom to conduct the research on my own, but would also always help me with scientific problems, be it acquiring new techniques, programming difficulties or – most importantly – he would *always* find a way out if I had reached an impasse where I would almost abandon hope. I truly have to thank him where I stand today. Thank you for always being encouraging and supportive over all these years. I could not imagine a better supervisor.

I would also like to thank my thesis advisory committee, Michael Pecka and Benedikt Grothe, who would always kindle many novel ideas, provide critical comments and help me out with theoretical problems whenever I needed it. My thanks also goes to Mike Myoga for many wonderful discussions that helped me understand things, that I didn't even know that I didn't understand. Financially, I was supported by the Sonderforschungsbereich 870 (SFB 870). Thank you also, of course, to the Graduate School of Systemic Neuroscience (GSN) with the absolutely wonderful and incredibly helpful people from the administration who I will truly miss and also all the other GSN students I met along the way, who not only made this graduate school into such an amazing environment to study in, but who also formed a great community that I enjoyed being part of and in which many friendships have arisen from.

Thank you to all the members of the Leibold lab, especially Olivia Lewis and Dustin Fetterhoff, for creating such a remarkable working atmosphere, for the recharging coffee breaks in the sun, for the many inspiring conversations and just by simply making work such an overall joyful experience. Also a big thank you to Dóra Csordás, my late night thesis buddy, whose constant motivation and cheerfulness could make some rather hard times of writing go by *swift*ly .

I want to thank all my great friends for making my PhD experience more than worthwhile. Especially, I would like to thank Andreas Hutzler and Urte Albers, Atanas Georgiev, Mario Breitner, Philipp Bruns, and of course Gerardo Pérez, for always being supportive, providing great life advice and just by simply being the most reliable and amazing friends one could imagine.

Finally, I would like to thank my family for their never-ending patience and support I have received from them my whole life. For being there in times of sadness and in times of joy. For helping me when I would doubt myself and for always encouraging me to go on no matter what it takes. For shaping me into the person I am today. For their endless love. I truly could not have done it without you. Thank you.

PUBLICATION LIST

ARTICLES

Groß, Sebastian, and Christian Leibold. 'A novel mammalian ITD encoding mechanism for sound localization in the azimuthal plane and sound source separation in cocktail party settings'. *In preparation*.

SELECT CONFERENCE PRESENTATIONS

Groß, Sebastian, and Christian Leibold (2019). 'Short-term ITD (interaural time difference) estimation of natural sound stimuli via effective models of binaural brainstem nuclei'. Abstract and poster presentation at conference *Neurowissenschaftliche Gesellschaft Gttingen*.

Groß, Sebastian, and Christian Leibold (2020). 'Sound source location in the azimuthal plane: separating sounds via short-term interaural time difference estimations'. Abstract and poster presentation at conference *Bernstein Conference Berlin*.

FIGURE PERMISSIONS

The following figures have been reprinted with the indicated permissions.

FIGURE 1.3

Reprinted with permission from Physiological Reviews, *Mechanisms of sound localization in mammals*, Benedikt Grothe, Michael Pecka, David McAlpine, VOL 90, July 2010, p. 995.

FIGURE 1.4

Reprinted with permission from Physiological Reviews, *Mechanisms of sound localization in mammals*, Benedikt Grothe, Michael Pecka, David McAlpine, VOL 90, July 2010, p. 991.

EIDESSTATTLICHE VERSICHERUNG / AFFIDAVIT

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation *Sound Source Coding in the Azimuthal Plane: Separating Sounds via Short-Term Interaural Time Difference Estimations* selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation *Sound Source Coding in the Azimuthal Plane: Separating Sounds via Short-Term Interaural Time Difference Estimations* is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, am 15. Juni, 2021
Munich, June 15, 2021

Sebastian Groß
Doctorand

AUTHOR CONTRIBUTIONS

The contributions of the authors Sebastian Groß (SG) and Christian Leibold (CL) to the modeling study conducted during my PhD are as follows:

CL conceived the modeling study. SG and CL designed the model. SG implemented the model. SG performed the modeling study with the help of CL. SG analyzed the data and interpreted the results with help of CL. SG designed the figures with the help of CL. SG wrote the manuscript.

We assert that aforementioned author contributions are correct and accurate:

Sebastian Groß	Prof. Dr. Christian Leibold
Doctorand	Supervisor